DOCUMENT RESUME

ED 210 314 TM 810 969

AUTHOR Vale, C. David; And Others
TITLE Methods for Linking Item Parameters. Final Report.
INSTITUTION Assessment Systems Corp., St. Paul, Minn.
SPONS AGENCY Air Force Human Resources Lab., Brooks AFB, Texas.
REPORT NO AFHRL-TR-81-10
PUB DATE Aug 81
CONTRACT F33615-80-C-0008
NOTE 190p.

EDRS PRICE MF01/PC08 Plus Postage.
DESCRIPTORS *Aptitude Tests; Armed Forces; Bayesian Statistics; *Item Analysis; Latent Trait Theory; Maximum Likelihood Statistics; Research Design; Simulation; Test Items; *Test Theory
IDENTIFIERS Adaptive Testing; *Item Calibration; Item Linking; *Parameter Identification

ABSTRACT

A simulation study to determine appropriate linking methods for adaptive testing items was designed. Three basic data sets for responses were created. These were randomly sampled, systematically sampled, and selected data sets. The evaluative criteria used were fidelity of parameter estimation, asymptotic ability estimates, root-mean-square error of estimates, and the correlation between true and estimated ability. Test length appeared more important to calibration effectiveness than sample size. Efficiency analyses suggested that increases in test length were several times as effective in improving calibration efficiency as proportionate increases in calibration sample sizes. The asymptotic ability analyses suggested that the linking procedures based on Bayesian ability estimation were more effective. The equivalent-tests method was no better than not linking. Bayesian scoring procedures were slightly superior to the others tested. Efficiency loss due to linking error was less than that due to item calibration error. Test length and sample size had a definite effect on calibration efficiency but no strong effects appear with respect to linking efficiency. For the systematically sampled data set, the anchor-test method produced the most efficient item pools in terms of linking efficiency. Bayesian scoring was preferred over the maximum likelihood scoring procedure. (Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

AFHRL-TR-81-10

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

AIR FORCE HUMAN RESOURCES LABORATORY

ED210314

METHODS FOR LINKING ITEM PARAMETERS

By

C. David Vale
Vincent A. Maurelli
Kathleen A. Gialluca
David J. Weiss
Assessment Systems Corporation
2395 University Avenue, Suite 306
St. Paul, Minnesota 55114

Malcolm James Ree

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
AFHRL
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

August 1981

Final Report

Approved for public release; distribution unlimited.

TM 810 969

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Public Affairs Office has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nationals.

This report has been reviewed and is approved for publication.

MALCOLM JAMES REE
Contract Monitor

NANCY GUINN, Technical Director
Manpower and Personnel Division

RONALD W. TERRY, Colonel, USAF
Commander

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFHRL-TR-81-10	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) METHODS FOR LINKING ITEM PARAMETERS		5. TYPE OF REPORT & PERIOD COVERED Final 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) C. David Vale Vincent A. Maurelli Kathleen A. Gialluca David J. Weiss Malcolm James Ree		8. CONTRACT OR GRANT NUMBER(s) F33615-80-C-0008
9. PERFORMING ORGANIZATION NAME AND ADDRESS Assessment Systems Corporation 2395 University Avenue, Suite 306 St. Paul, Minnesota 55114		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61101F ILIR0018
11. CONTROLLING OFFICE NAME AND ADDRESS HQ Air Force Human Resources Laboratory (AFSC) Brooks Air Force Base, Texas 78235		12. REPORT DATE August 1981 13. NUMBER OF PAGES 190
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Manpower and Personnel Division Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) adaptive testing Bayesian item analysis item linking latent trait theory logistic model maximum likelihood parameter estimation test equating		

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A simulation study to determine appropriate linking methods for adaptive testing items was designed. Responses of examinees of three group sizes for four test lengths were simulated. Three basic data sets were created: (a) randomly sampled data set, (b) systematically sampled data set, and (c) selected data set. Three categories of evaluative criteria were used: fidelity of parameter estimation, asymptotic ability estimates, root-mean-square error of estimates, and the correlation between true and estimated ability. Test length appeared to be relatively more important to calibration effectiveness than was sample size, efficiency analyses suggested that increases in test length were at least three to four times as effective in improving calibration efficiency as proportionate increases in calibration sample sizes. The asymptotic ability analyses suggested that the linking procedures based on Bayesian

Item 20 (Continued):

ability estimation (an equivalent-groups procedure) were somewhat more effective than the others and that the equivalent-tests method was typically no better than not linking at all. Analyses using the relative efficiency criteria suggested that the equivalent-groups procedures were superior to the equivalent-tests procedures and that those using Bayesian scoring procedures were slightly superior to the others tested. Efficiency loss due to linking error was always less than that due to item calibration error and although test length and sample size had a definite effect on calibration efficiency, no strong effects appear with respect to linking efficiency. For the systematically sampled data set, the anchor-test and anchor-group methods were considered along with the equivalence methods. In terms of linking efficiency, the anchor-test method produced the most efficient item pools. The anchor-group method resulted in efficiencies equivalent to those of the anchor-test procedure if large groups were used, but with smaller groups the efficiencies dropped somewhat. The equivalence methods were somewhat less efficient than either of the anchor methods. Bayesian scoring was preferred over the maximum-likelihood scoring procedure. An application of the results of this research to a practical linking problem was described with equivalent-groups linking. An anchor-test linking method was suggested for adding items at later times.

SUMMARY

Objective

The objective was to determine appropriate methods for linking parameters of test items under a variety of testing conditions.

Background

Computerized adaptive testing (CAT) is a form of test administration that the Armed Services may soon implement. It requires that large numbers of items be calibrated and stored in item banks from which specific items are drawn adaptively by the computers for each testee. Because the number of items to be calibrated is so large, it is not feasible to administer all of them to a single group, and so the items must be calibrated in separate sets and then linked together onto a common scale. Four different methods of linking the item sets were devised and evaluated.

Approach

In an evaluation of the adequacy of various linking methods, the true item parameters must be known. These were obtained through a computer simulation study with a design based on a practical testing environment.

Specifics

Method. A simulation study was designed in which simulated test items were defined to be similar in terms of their item parameters to Armed Services test items, and populations of simulated examinees were defined to be similar in ability to those individuals likely to take Armed Services tests.

Four linking methods were evaluated. The *equivalent-groups method* linked items by assuming examinee groups to be equivalent. The *equivalent-tests method* assumed tests to contain equivalent items. The *anchor-group method* linked through a common group of examinees. The *anchor-test method* linked through a common set of items. These methods were compared to each other and to a condition in which no explicit linking was done.

Three linking conditions were simulated. One was the condition in which test booklets were randomly distributed among the entire population. Another was the condition in which test booklets were distributed systematically among relatively few testing centers. The final condition was one in which a population of examinees selected on the basis of their scores was used.

Three categories of evaluative criteria were used. Fidelity-of-parameter-estimation criteria examined the relations between true and estimated item parameters. Asymptotic-ability-estimate criteria examined the relations between the true and asymptotic (i.e., infinite-test-length) ability estimates. Efficiency-of-ability-estimation criteria included average item information and relative efficiency.

Findings and discussion. Despite its simplicity, the *equivalent-groups method* worked well under most testing conditions. The *anchor-group* and *anchor-test methods* were slightly superior when the assumption of equivalent groups was violated. The *equivalent-tests method* was generally less effective than the other three methods. Modal-Bayesian scoring of tests generally produced better linking results than did maximum-likelihood scoring.

Conclusions

Two procedures can be recommended for linking. Linking during development of the initial item pool can most efficiently be accomplished using the *equivalent-groups method*, with examinees randomly selected from the general calibration population. Items added to the pool at a later date should be linked using the *anchor-test method*.

PREFACE

This effort was carried out under ILIR0018, Methods for Linking Item Parameters. It was basic research conducted in support of an ongoing program in the area of Assessment of Personnel Qualification which supports the general thrust area of Manpower and Force Management. It was performed to gain knowledge in advanced psychometric theory as applied to computer driven adaptive testing, item banking, and Item Response Theory. This report is one in a series aimed at advancing the state of the art in the measurement of human characteristics

The authors wish to thank James B. Sympson for his suggestions and insightful criticism of portions of an earlier draft of this report.

TABLE OF CONTENTS

Page

I. INTRODUCTION 13
    Overview of Item Response Theory. 13
    Item Calibration. 16
        Estimation Techniques. 16
            Maximum-likelihood estimation 16
            Minimum chi-square estimation 17
        Criteria of Good Estimation. 17
        Evaluation of Estimation Techniques. 18
    Item Linking. 22
        Predicting, Equating, and Linking--A Clarification of Concepts 22
            Equating and predicting 23
            Linking 23
        Paradigms of Linking and Equating. 24
            Methods based on sampling 25
            Methods based on anchoring. 26
            Composite network methods 27
        Criteria of Linking Adequacy 27
        Evaluation of Linking Techniques 28
            Rasch model 28
            Three-parameter logistic model. 33
    Conclusions 38

II. BASIC RESEARCH DESIGN. 40
    Development of Simulation Models. 41
        Specification of Items 41
            Analyses of ASVAB item parameters 41
            Specification of a representative item domain 46
        Specification of Ability Distributions 49
            Examinee data available 49
            Score data available. 49
            Raw score analysis. 50
            Differences among AFEES 51
            Modal Bayesian trait estimates. 56
            Specification of distributional parameters. . 57
        Basic Data Sets. 58
            Randomly sampled examinees. 58
            Systematically sampled examinees. 58
            Selected examinees. 59
            Composite sets of items 59
            Calibration of items. 59
    Evaluative Criteria 60
        Fidelity of Parameter Estimation 61
            Bias. 61
            Absolute error. 61
            Root-mean-square error. 61

Page

Correlations 61
Characteristics of Asymptotic Ability Estimates . . 61
Mean and standard deviation 63
Absolute and root-mean-square error 63
Correlation 63
Efficiency of Ability Estimation 64
Information 65
Relative efficiency 66

III. EVALUATION OF THE BASIC DATA SETS 67
Randomly Sampled Examinees 67
Fidelity of Parameter Estimation 67
Characteristics of Asymptotic Ability Estimates . . 71
Efficiency of Ability Estimation 72
Systematically Sampled Examinees 74
Fidelity of Parameter Estimation 74
Characteristics of Asymptotic Ability Estimates . . 78
Efficiency of Ability Estimation 79
Selected Examinees 79
Fidelity of Parameter Estimation 79
Characteristics of Asymptotic Ability Estimates . . 82
Efficiency of Ability Estimation 82
Conclusions 83

IV. LINKING WHEN EXAMINEES ARE RANDOMLY SAMPLED 85
Equivalence Methods 85
Procedure 85
Equivalent groups 85
Equivalent tests 89
Results 89
Fidelity of parameter estimation 90
Characteristics of asymptotic ability estimates 93
Efficiency of ability estimation 96
Discussion 100
Conclusions 100

V. LINKING WHEN EXAMINEES ARE SYSTEMATICALLY SAMPLED 102
Equivalence Methods 102
Procedure 102
Results 102
Fidelity of parameter estimation 102
Characteristics of asymptotic ability estimates 106
Efficiency of ability estimation 108
Discussion 111
Anchor-Group Method 112
Procedure 112

Page

Results--Modal Bayesian Scores 112
Fidelity of parameter estimation. 112
Characteristics of asymptotic ability estimates 116
Efficiency of ability estimation. 118
Results--Robust-Maximum-Likelihood Scores. 119
Fidelity of parameter estimation. 119
Characteristics of asymptotic ability estimates 121
Efficiency of ability estimation. 122
Discussion 123
Anchor-Test Method. 124
Procedure. 124
Generation of the source item pool. 124
Selection of anchor-test items. 124
Determination of the linking transformations. 126
Results--Modal Bayesian Scores 126
Fidelity of parameter estimation. 126
Characteristics of asymptotic ability estimates 132
Efficiency of ability estimation. 135
Results--Robust-Maximum-Likelihood Scores. 137
Fidelity of parameter estimation. 137
Characteristics of asymptotic ability estimates 139
Efficiency of ability estimation. 141
Discussion 141
Conclusions 143

VI. LINKING WHEN EXAMINEES ARE SELECTED. 145
Equivalence Methods 145
Procedure. 145
Results. 145
Fidelity of parameter estimation. 145
Characteristics of asymptotic ability estimates 146
Efficiency of ability estimation. 147
Anchor-Group Method 148
Procedure. 148
Results. 148
Fidelity of parameter estimation. 148
Characteristics of asymptotic ability estimates 150
Efficiency of ability estimation. 151
Discussion 151
Anchor-Test Method. 153
Procedure. 153
Results. 153

Page

Fidelity of parameter estimation. 153
Characteristics of asymptotic ability estimates 155
Efficiency of ability estimation. 156
Discussion 158
Conclusions 158

VII. PRACTICAL APPLICATIONS OF LINKING. 160
Development of a Composite Approach 160
Linking the Initial Item Set--A Summary of Findings 160
Linking Across Time--Further Analyses. 161
Method. 162
Results 162
Discussion. 165
Design for a Specific Application 166
Description of the Problem 166
A Proposed Linking Design. 167

VIII. SUMMARY AND CONCLUSIONS. 169
Summary 169
Previous Literature. 169
Linking Criteria 169
Simulation Design. 170
Results. 171
Application to a Practical Linking Problem 172
Conclusions 172

REFERENCES. 175

APPENDICES

Appendix A: Supporting Tables. 181
Appendix B: Revisions to Program OGIVIA. 185

LIST OF ILLUSTRATIONS

Figure Page

1 Item Characteristic Curves. 15

2 Sympson's Data Collection Plan. 38

3 Raw Score Frequency Distribution--Word Knowledge. 51

4 Raw Score Frequency Distribution--Arithmetic Reasoning. . . . 52

5 Raw Score Frequency Distribution--Mathematics Knowledge . . . 52

6 Raw Score Frequency Distribution--Electronics Information . . 53

7 Raw Score Frequency Distribution--Mechanical Comprehension. . 53

8 Raw Score Frequency Distribution--General Science 54

9 True Information Curves, Using True Item Parameters, for Each of Three Anchor Tests. 125

LIST OF TABLES

Table Page

1 Number of Items in the Two Sets of Item Parameter Data. . . . 42

2 Item Parameter Summary Statistics from Experimental Form 8 and New Forms 8, 9, 10 43

3 Numbers and Percentages of Items From the New Forms 8, 9, 10 With _a_ Parameters Set Equal to the Maximum Value 45

4 Numbers and Percentages of Items From Experimental Form 8 With _a_ Parameters Equal to or Exceeding 2.40. 46

5 Parameter Intercorrelations for Experimental Form 8 and New Forms 8, 9, 10. 47

6 Overall Skew and Kurtosis--ASVAB-7 Number-Correct Scores (N=32,444) 50

7 Standard-Score Summary Statistics Across AFEES for ASVAB-7 Subtests. 55

8 Mean, Standard Deviation, Skew, and Kurtosis of ASVAB-8 Modal Bayesian Ability Estimates (N=500) 56

9 Item Parameter Bias--Basic Data Set--Randomly Sampled Examinees. 68

10 Parameter Correlations--Basic Data Set--Randomly Sampled Examinees. 69

11 Absolute Parameter Error--Basic Data Set--Randomly Sampled Examinees. 70

12 Root-Mean-Square Parameter Error--Basic Data Set--Randomly Sampled Examinees 71

13 Absolute Asymptotic Ability Error--Basic Data Set--Randomly Sampled Examinees 72

14 Root-Mean-Square Asymptotic Ability Error--Basic Data Set--Randomly Sampled Examinees 72

15 Mean Relative Efficiency--Basic Data Set--Randomly Sampled Examinees. 73

16 Item Parameter Bias--Basic Data Set--Systematically Sampled Examinees. 74

Table Page

17 Parameter Correlations--Basic Data Set--Systematically Sampled Examinees 75

18 Absolute Parameter Error--Basic Data Set--Systematically Sampled Examinees 76

19 Root-Mean-Square Parameter Error--Basic Data Set--Systematically Sampled Examinees 77

20 Absolute Asymptotic Ability Error--Basic Data Set--Systematically Sampled Examinees 78

21 Root-Mean-Square Asymptotic Ability Error--Basic Data Set--Systematically Sampled Examinees 78

22 Relative Efficiency--Basic Data Set--Systematically Sampled Examinees. 79

23 Item Parameter Bias--Basic Data Set--Selected Examinees. . . 80

24 Parameter Correlations--Basic Data Set--Selected Examinees . 80

25 Absolute Parameter Error--Basic Data Set--Selected Examinees. 81

26 Root-Mean-Square Parameter Error--Basic Data Set--Selected Examinees 82

27 Asymptotic Ability Error--Basic Data Set--Selected Examinees. 82

28 Relative Efficiency--Basic Data Set--Selected Examinees. . . 83

29 Item Parameter Error--Equivalence Methods--Homogeneous Condition Using Randomly Sampled Examinees 90

30 Item Parameter Error--Equivalence Methods--Heterogeneous Condition Using Randomly Sampled Examinees 93

31 Asymptotic Ability Estimates--Equivalence Methods--Homogeneous Condition Using Randomly Sampled Examinees . . 94

32 Asymptotic Ability Estimates--Equivalence Methods--Heterogeneous Condition Using Randomly Sampled Examinees . 95

33 Efficiency Analysis--Equivalence Methods--Homogeneous Condition Using Randomly Sampled Examinees 96

Table Page

34 Efficiency Analysis--Equivalence Methods--Heterogeneous Condition Using Randomly Sampled Examinees 98

35 Cellwise Efficiency Analysis--Bayesian Score--Randomly Sampled Examinees 99

36 Cellwise Efficiency Analysis--Equivalent Tests--Randomly Sampled Examinees 99

37 Item Parameter Error--Equivalence Methods--Homogeneous Condition Using Systematically Sampled Examinees 103

38 Item Parameter Error--Equivalence Methods--Heterogeneous Condition Using Systematically Sampled Examinees 105

39 Asymptotic Ability Estimates--Equivalence Methods--Homogeneous Condition Using Systematically Sampled Examinees. 107

40 Asymptotic Ability Estimates--Equivalence Methods--Heterogeneous Condition Using Systematically Sampled Examinees. 108

41 Efficiency Analysis--Equivalence Methods--Homogeneous Condition Using Systematically Sampled Examinees 109

42 Efficiency Analysis--Equivalence Methods--Heterogeneous Condition Using Systematically Sampled Examinees 110

43 Cellwise Efficiency Analysis--Bayesian Score--Systematically Sampled Examinees 110

44 Cellwise Efficiency Analysis--Equivalent Tests--Systematically Sampled Examinees 111

45 Item Parameter Error--Anchor Groups--Homogeneous Condition Using Systematically Sampled Examinees 113

46 Item Parameter Error--Anchor Groups--Heterogeneous Condition Using Systematically Sampled Examinees 115

47 Asymptotic Ability Estimates--Anchor Groups--Homogeneous Condition Using Systematically Sampled Examinees 116

48 Asymptotic Ability Estimates--Anchor Groups--Heterogeneous Condition Using Systematically Sampled Examinees 117

Table Page

49 Efficiency Analysis--Anchor Groups--Homogeneous Condition Using Systematically Sampled Examinees 118

50 Efficiency Analysis--Anchor Groups--Heterogeneous Condition Using Systematically Sampled Examinees 119

51 Item Parameter Error--Anchor Groups--Homogeneous Condition Using Systematically Sampled Examinees 120

52 Asymptotic Ability Estimates--Anchor Groups--Homogeneous Condition Using Systematically Sampled Examinees 122

53 Efficiency Analysis--Anchor Groups--Homogeneous Condition Using Systematically Sampled Examinees 123

54 Item Parameter Error--Anchor Tests--Homogeneous Condition Using Systematically Sampled Examinees 127

55 Item Parameter Error--Anchor Tests--Heterogeneous Condition Using Systematically Sampled Examinees 130

56 Asymptotic Ability Estimates--Anchor Tests--Homogeneous Condition Using Systematically Sampled Examinees 133

57 Asymptotic Ability Estimates--Anchor Tests--Heterogeneous Condition Using Systematically Sampled Examinees 134

58 Efficiency Analysis--Anchor Tests--Homogeneous Condition Using Systematically Sampled Examinees 135

59 Efficiency Analysis--Anchor Tests--Heterogeneous Condition Using Systematically Sampled Examinees 136

60 Item Parameter Error--Anchor Tests--Homogeneous Condition Using Systematically Sampled Examinees 138

61 Asymptotic Ability Estimates--Anchor Tests--Heterogeneous Condition Using Systematically Sampled Examinees 140

62 Efficiency Analysis--Anchor Tests--Homogeneous Condition Using Systematically Sampled Examinees 142

63 Item Parameter Error--Equivalence Methods--Homogeneous Condition Using Selected Examinees 146

64 Asymptotic Ability Estimates--Equivalence Methods--Homogeneous Condition Using Selected Examinees 147

Table Page

65 Efficiency Analysis--Equivalence Methods--Homogeneous Condition Using Selected Examinees 148

66 Item Parameter Error--Anchor Groups--Homogeneous Condition Using Selected Examinees 149

67 Asymptotic Ability Estimates--Anchor Groups--Homogeneous Condition Using Selected Examinees 151

68 Efficiency Analysis--Anchor Groups--Homogeneous Condition Using Selected Examinees 152

69 Item Parameter Error--Anchor Tests--Homogeneous Condition Using Selected Examinees 154

70 Asymptotic Ability Estimates--Anchor Tests--Homogeneous Condition Using Selected Examinees 156

71 Efficiency Analysis--Anchor Tests--Homogeneous Condition Using Selected Examinees 157

72 Asymptotic Ability Metric of Cascaded Tests--Modal Bayesian Scoring 163

73 Asymptotic Ability Metric of Cascaded Tests--Maximum-Likelihood Scoring 164

74 Linkage Efficiency of Cascaded Tests--Modal Bayesian Scoring 165

75 Linkage Efficiency of Cascaded Tests--Maximum-Likelihood Scoring 165

I. INTRODUCTION

During the past decade, an extensive investigation of adaptive testing has been conducted. In its simplest form, adaptive testing amounts to administering the subset of items, selected from a larger pool, that provides the most information about the individual regarding the characteristic the test measures. A summary of the current state of the art, extracted from the 1979 Computerized Adaptive Testing Conference (Weiss, 1980), is that adaptive testing potentially offers several advantages over conventional testing methods, but to realize these advantages, characteristics of the items comprising the pool must be accurately determined.

Most adaptive testing technology is built on the framework of Item Response Theory (IRT), also called Latent Trait Theory or Item Characteristic Curve (ICC) Theory. In IRT, test items are described by a set of item parameters. It is these parameters that must be accurately determined if adaptive testing is to be effective. This determination is called item calibration. Because adaptive testing requires a large item pool, and because item calibration requires administration to a large number of examinees, calibration must often be accomplished in parts such that different groups of individuals take different sets of items.

The purposes of the project were to determine efficient methods of partitioning the calibration examinee samples and item sets, and to determine efficient methods of re-assembling or linking the parts into a common whole once the individual calibrations are accomplished. As background to the research, the first section of this report reviews some of the concepts basic to calibration and linking. Previous research, its shortcomings and unanswered questions, will be reviewed and discussed. In subsequent sections, a research design to eliminate these shortcomings will be described and research conducted according to that design will be reported.

Overview of Item Response Theory

Item Response Theory has been called the psychometric equivalent of Einstein's Theory of Relativity (Warm, 1978). Stated simply, IRT specifies a general mathematical relationship between an individual's status on an underlying trait, characteristics of a test item, and the probabilities regarding how the individual will respond to the item. The term IRT actually refers to a general class of psychometric models. Included in the class are models for use when the response is dichotomous (Lord & Novick, 1968; Birnbaum, 1968), models for use when the response is polychotomous (Samejima, 1969, 1972; Bock, 1972), and models for use when the response is continuous (Samejima, 1974). These models have typically been developed for use where a unidimensional trait is measured. Extension of each to multidimensional traits would double the number of available models. Hambleton & Cook (1977) present an overview of most of the unidimensional IRT models.

All the item domains considered by the current research contained dichotomous ability items of a multiple-choice nature. Two IRT models are appropriate for such items: the three-parameter normal and logistic ogive models. For reasons of mathematical tractability, the logistic model is generally preferred over the normal model and will be of primary focus throughout this report. A single-parameter degenerate case of the three-parameter logistic model, the Rasch model, will be included in some parts of this review because of its similarity to the three-parameter logistic model and because more research has been done on calibration and linking using the Rasch model than has been done using the three-parameter logistic model.

In the three-parameter logistic model, the item is characterized by the three parameters a, b, and c. Ability is characterized by a single parameter, theta. The a parameter is an index of the item's power to discriminate among different levels of ability. It ranges, theoretically, between negative and positive infinity but practically between zero and about three when ability is expressed in a standard-score metric. A negative a parameter would mean that a low-ability examinee had a better chance of answering the item correctly than did a high-ability examinee. An a parameter of zero would mean that the item had no capacity to discriminate between different levels of ability (and would therefore be useless as an item in a power test). Items with high positive a parameters provide sharper discrimination among levels of ability and are generally more desirable than items with low a parameters.

The b parameter indicates the difficulty level of an item. It is scaled in the same metric as ability and indicates the value of theta an examinee would need in order to have a 50-50 chance of knowing the correct answer to the item. This is not, however, the level of theta at which the examinee has a 50-50 chance of selecting a correct answer if it is possible to answer the item correctly by guessing.

The c parameter gives the probability with which a very low-ability examinee would answer the item correctly. It is often called the guessing parameter as it is roughly the probability of answering the item correctly if the examinee does not know the answer and guesses at random. Intuitively, the c parameter of an item should be the reciprocal of the number of alternatives in the item. Empirically, it is typically somewhat lower than this.

All four parameters enter into the three-parameter logistic test model to determine the probability of a correct response. The formal mathematical relationship is given by Equation 1:

$$P(u=1|\theta) = c + (1-c)\Psi[1.7a(\theta-b)] \qquad [1]$$

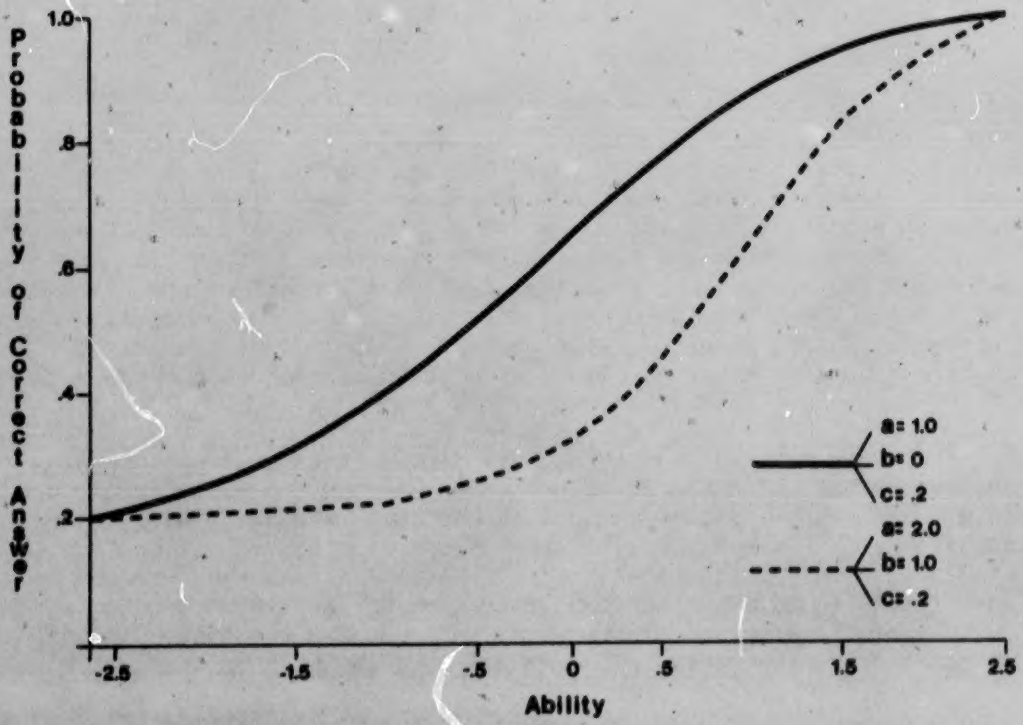
where:

$$\Psi(x) = [1+\exp(-x)]^{-1}$$

In Equation 1, $u = 1$ if the response to the item is correct and $u = 0$ if the response is incorrect. The relationship expressed in Equation 1 is shown graphically in Figure 1. The item characteristic curve drawn with a solid line is for an item with $a = 1.0$, $b = 0.0$, and $c = .2$. The slope at any point is related to a. The lower asymptote corresponds to a probability or c of .2. The item characteristic curve shown with a dashed line is for an item with $a = 2.0$, $b = 1.0$, and $c = .2$. The midpoint of the curve has shifted to $\theta = 1.0$. The slope of the curve is steeper near $\theta = b$. The lower asymptote of the curve remains, however, at .2.

Ultimately, theta is the only parameter that needs to be estimated; the objective of testing is to estimate an individual's ability level. To accomplish this, however, it is necessary to first know the item parameters. The items must therefore be calibrated.

Figure 1. Item Characteristic Curves



Although Ree (1979) has shown that, under certain conditions, ability estimation can proceed very well with quite poor estimates of item parameters, in the general case, good estimation of ability requires good estimation of item parameters.

Item Calibration

Estimation Techniques

Two methods of estimating item parameters have been primarily employed in IRT applications: maximum-likelihood estimation and minimum chi-square estimation. The former method identifies the parameter values for which the probability of observing the observed data (i.e., the likelihood) is a maximum. The latter method identifies the parameter values for which the discrepancy between the model and the observed data is a minimum. Both methods are discussed in detail below with general reference to three-parameter models.

Maximum-likelihood estimation. Conceptually, the application of maximum-likelihood techniques to estimation of item parameters is simple. The probability of observing a response vector is expressed in terms of the unknown parameters, and the parameter values making this probability a maximum are the maximum-likelihood parameter estimates. In practical calibration applications, however, the number of parameters to be estimated may exceed several thousand and the numerical difficulties make the simple conceptual task practically formidable.

Two approaches to maximum-likelihood item calibration are the unconditional and the conditional approaches[1] (Bock, 1972; Bock & Lieberman, 1970). In the unconditional approach, a distribution of theta is assumed and the theta parameter in each individual response vector is integrated out. This results in a set of likelihood functions, one function for each examinee, that is independent of theta. From these functions, the item parameters can be estimated. There are two difficulties with use of the unconditional approach. First, it requires an assumption as to the form of the distribution of theta and, second, due to the integration required, it is computationally too burdensome for use with more than a few items.

---

[1] The terms "unconditional" and "conditional" as used here should not be confused with the identical terms used in the Rasch literature (e.g., Anderson, 1971, 1977; Gustafsson, 1979; Reckase, 1977). "Unconditional" in the Rasch literature refers to the "conditional" case discussed here. "Conditional" in the Rasch literature refers to the use of likelihood functions conditioned on the sufficient number correct statistic and is, in some ways, analogous to the "unconditional" approach discussed here.

The conditional approach assumes the theta values are unknown but fixed parameters to be estimated in the same manner as the item parameters. The computer program LOGIST (Wood, Wingersky, & Lord, 1976) is the major operationalization of the conditional approach to calibration. Although, in theory, both theta parameters and item parameters can be estimated simultaneously, LOGIST iterates between estimation of theta and estimation of item parameters. Provisional values of theta are obtained from each examinee's raw score and these are used as true theta values while the item parameters are estimated. The estimated item parameters are then used to re-estimate the theta parameters and the procedure iterates until stable item and theta parameter estimates are found. Convergence can require a large amount of computation.

Minimum chi-square estimation. Regardless of how the parameters of the model are estimated, the adequacy with which the model fits the observed data can be tested with a Pearson chi-square test. This is accomplished by grouping subjects on the basis of ability (or estimated ability), predicting for each item the proportion of subjects in each subgroup who should answer it correctly according to the model, and testing the significance of the discrepancy between observed and predicted proportions using a chi-square test. The minimum chi-square approach to estimation explicitly selects parameter values to minimize this chi-square value. Except for the change in criterion, however, the approach is similar to the conditional maximum-likelihood approach.

A major proponent of this approach was Urry (1978), who sponsored several computer programs to perform such estimation; the most frequently used are OGIVIA and ANCILLES. In these programs, examinees are scored based on provisional parameter estimates. Several trial values of the $\underline{c}$ parameters are chosen and $\underline{a}$ and $\underline{b}$ parameters are estimated using equations given by Urry (1976). The combination of $\underline{a}$, $\underline{b}$, and $\underline{c}$ that produces the minimum lack of fit with the IRT item characteristic curve, as indicated by a chi-square statistic, is chosen as the minimum chi-square parameter estimate.

Criteria of Good Estimation

Texts in statistics (e.g., Lindgren, 1976) typically list four desirable characteristics of an estimator of a parameter: an estimator should be unbiased, efficient, sufficient, and consistent. An unbiased estimator has an expected value equal to the parameter it estimates. An efficient estimator has, in comparison to other estimators, small mean squared-deviation from the parameter. If the estimator is unbiased, its variance is an index of its efficiency. A sufficient estimator contains all the information regarding the parameter that is available from the data on which it is calculated. Information of an unbiased estimator is an estimate of the reciprocal of the squared error of estimate of the parameter (see Lindgren, 1976, for a discussion of information). An unbiased sufficient estimator is efficient in an absolute sense as no other estimator can be more efficient. Finally, a consistent estimator is one that converges on the parameter values as the data on which it is based increase. Increased data, in psychometric applications, refers to both increased subject sample size and increased item set size (i.e., more items). Both must approach infinity for item and ability parameter estimates to converge on their true values, but acceptable estimates can be obtained from sample sizes that are obtainable in practice.

Evaluation of the quality of estimators in terms of these criteria can be done analytically in simple applications. In evaluation of item calibration techniques, analytic calculation of these criteria is practically impossible because of the complexity of the calculations. Hence, they must be evaluated through simulation techniques. In such a simulation, responses to items with known parameters are generated according to a statistical model (see Vale & Weiss, 1975, or Ree, 1978, for a full description of a simulation). Parameters are then estimated from the item responses as if these responses had been generated by real examinees, and the estimated parameters are compared to the true values. In studies done comparing estimated with true item parameters, three indices of comparison have typically been calculated for individual item parameters. The average algebraic difference between true and estimated parameters has been calculated as an index of bias. The mean-square deviation of estimated parameters from the true parameters has been calculated and can be considered an index of efficiency. The correlation between true and estimated parameter values has been calculated and, if the estimates are linear estimates of the parameters, this can be thought of as an index of relative sufficiency when comparing two methods on the same items and subjects. All these indices are typically calculated at several combinations of test length and sample size and thus provide some evidence for consistency.

In addition to evaluation of the parameters separately, some researchers (e.g., Ree, 1978) have attempted to evaluate the parameters collectively by comparing the test scores produced by the estimated parameters with those produced by the true parameters. There may be some tendency for errors in one parameter to cancel out or compensate for errors in other parameters. Separate evaluation would not show this effect; joint evaluation would. As will be discussed in regard to the study by Ree, this evaluation may be done in several ways.

Evaluation of Estimation Techniques

Lord (1975) evaluated the LOGIST procedure in a simulation study. For this study, item parameters for 90 verbal items of the Scholastic

Aptitude Test were estimated by LOGIST using a sample of 2,995 examinees. These parameters, after correction for errors of estimate, were used as the basis for a Monte-Carlo simulation in which 2,995 hypothetical examinees (with abilities similar to those of real examinees) "responded" to the items according to the logistic test model. These responses were then used by LOGIST to re-estimate the item parameters. The parameters entering the simulation model were taken to be true parameters, and the effectiveness of LOGIST was evaluated by how accurately these true parameters were estimated. Root-mean-square errors of estimation and the correlations between true and estimated parameters were, respectively, .130 and .920 for the a parameters and .196 and .988 for the b parameters. For the c parameters, the root-mean-square error was .070; the correlation between the true and estimated c parameters was not reported.

Gugel, Schmidt, and Urry (1976) reported a similar simulation study of the minimum chi-square procedure. Some major differences between this study and that of Lord's (in addition to the different estimation procedure) were that (a) the hypothetical subjects were drawn from a standard normal ability distribution rather than matched to subjects having taken an existing test, (b) the hypothetical item parameters were rectangularly distributed in ranges typical for such parameters rather than matched to those from an existing test, and (c) subject sample sizes and item set sizes were systematically varied. Of the conditions investigated a condition with 90 items and 2,000 subjects was most comparable to Lord's study of LOGIST. In this condition, root-mean-square errors and correlations were, respectively, .244 and .871 for the a parameter, .149 and .996 for the b parameter, and .069 and .568 for the c parameter. Direct comparisons with Lord's study are not particularly meaningful, however, because the distributions of all parameters were different and this can drastically affect the comparative indices. The study did note, however, that the minimum chi-square procedure did not work well when the numbers of subjects used fell as low as 500.

Schmidt and Gugel (1976) again reported the preceding study, as well as a second study in which the number of items used was 100 and the sample sizes were 2,000 and 3,000. Root-mean-square errors for the final estimates at sample sizes of 2,000 and 3,000, respectively, were .242 and .228 for the a parameter, .123 and .148 for the b parameter, and .056 in both samples for the c parameter. Correlations were .915 and .918 for the a parameter, .996 and .997 for the b parameter, and .764 and .760 for the c parameter. Little change was apparent between sample sizes of 2,000 and 3,000. The results of these two studies led Schmidt and Gugel to conclude that, as a rule-of-thumb, item sets should contain at least 100 items and should be administered to at least 2,000 subjects to obtain an accurate calibration.

Two studies comparing different calibration techniques have been done, to date. Ree (1978, 1979) compared four calibration techniques in three different populations. The four calibration techniques were: (a) ANCILLES, minimum chi-square estimation with ancillary correction for errors in estimation of ability, (b) OGIVIA, minimum chi-square estimation similar to that of ANCILLES, (c) LOGIST, the conditional maximum likelihood approach, and (d) transformation of classical parameters derived from IRT assuming a normal distribution of ability (see Jensema, 1976, for a description of the transformations). The three ability distributions were: (a) a rectangular distribution of ability bounded at $\theta = \pm 2.5$, (b) a normal (0,1) distribution of ability with elimination of the lower third on the basis of a number correct score, and (c) a normal (0,1) distribution of ability. The hypothetical items used in the simulation had parameters distributed normally in ranges typically found in real item sets. Among the criteria investigated were: (a) correlations between true and estimated item parameters, (b) correlations between ability estimates computed using both true and estimated item parameters, (c) correlations between true number-correct scores generated using both true and estimated item parameters, and (d) test information curves resulting from the true and estimated item parameters. All analyses were performed on samples of 2,000 examinees and tests 80 items in length.

Evaluated on the criterion of correlation between estimated and true item parameters, LOGIST generally produced the highest correlations. The exception to this was in the normal ability distribution in which OGIVIA produced slightly better estimates of a and b. The best estimates of the item parameters were obtained using LOGIST and a rectangular distribution of ability.

Correlations between true and estimated ability levels showed LOGIST to be slightly better than ANCILLES and OGIVIA, and the transformations to be slightly worse. Differences among correlations were small, however, ranging from .955 to .974 in the rectangular distribution, from .930 to .943 in the truncated normal distribution, and from .961 to .965 in the normal distribution.

Correlations between true scores obtained using true and estimated parameters showed very little difference among methods and only a small deviation from unity. The largest difference observed was in the rectangular distribution where the transformation yielded a correlation of .9910 and LOGIST yielded one of .9960. All other distributions produced correlations of .999, with variations in the fourth decimal place.

When compared in terms of the information curves produced by the item parameter estimates, all methods except the transformations produced information curves similar to the true information curve in the rectangular and normal ability distributions. In both of these distributions, LOGIST produced information curves somewhat closer to the true curve than did ANCILLES or OGIVIA. In the selected distribution, all methods produced noticeable departures from the true information curve.

Of the four criteria investigated, only the correlations among item parameters and the information curves are independent of the ability distribution; thus, these criteria are the only ones that can be compared across ability distributions. (Equivalent estimation accuracy would yield differences in the other criteria solely as a function of the ability distribution.) On these two criteria, LOGIST was nearly always superior to the other methods. The degree of superiority was not overwhelming, however, and an analysis of cost suggested that other methods were to be favored. The second-best procedure, in terms of psychometric criteria, was OGIVIA. OGIVIA required less than one-tenth as much computer time to use as did LOGIST.

As a final point, the level of correlation between actual and estimated ability levels and actual and estimated true scores is noteworthy. Especially with the true scores, the level of correlation was so high as to suggest that one might do well enough without bothering to estimate parameters at all. In fact, Ree (1979) has shown that the correlation between the estimated and true values of any one of the three IRT parameters can be degraded to little relation with its true value and still yield correlations between actual and estimated true scores of .98 and above. All these results, however, were obtained using conventional tests where all examinees answer the same items. When administration is adaptive and each examinee answers a different set of items, these correlations could be expected to drop substantially as a result of poor item calibration. Unfortunately, no study has investigated this effect directly. Schmidt and Gugel (1976), in the study discussed earlier, provided data that hinted at the answer. When the size of the calibration sample fell to 1,000 examinees and the length of the calibration item set fell to 60, there was a noticeable decrease in the quality of tests administered using a Bayesian strategy when compared to similar tests given using true item parameter values. Thus, although definitive data do not exist, those data which do exist suggest that the extremely high correlations between estimates of true scores obtained using the different parameter estimates may be due to an averaging-out phenomemon peculiar to conventionally administered tests.

The second study comparing various calibration procedures was done by Swaminathan and Gifford (1980). Noting that the Ree study investigated only a single test length and sample size, they compared ANCILLES and LOGIST in simulation at test lengths of 10, 15, 20, and 80 items and sample sizes of 50, 200, and 1,000. Items had true $\underline{a}$ parameters distributed rectangularly between .6 and 2.0, true $\underline{b}$ parameters distributed rectangularly between -2.0 and 2.0, and true $\underline{c}$ parameters fixed at .25. Three distributions of ability were used;

one was normally distributed with a mean of zero and variance of one, the second was rectangularly distributed between -1.73 and 1.73, and the third was a standardized negatively skewed beta distribution. Criteria of calibration effectiveness included the differences between means of true and estimated a, b, and c parameters, the correlations between true and estimated a and b parameters, the differences in means of ability estimates using true and estimated parameters, and the correlations between these values.

The b parameter estimates correlated highly with their true values in all conditions using either of the calibration methods. Medians for each of the distributions were all above .9. A trend toward higher correlations with increased test length was observed, and median correlations for LOGIST were slightly higher than those for ANCILLES. No substantial differences were observed among distributions.

The a parameters were less well estimated. Median correlations were near .4 for the normal and rectangular ability distributions, but dropped to near .2 in the skewed distribution. Improvements in estimation occurred both with increasing test length and sample size, however. Median correlations using LOGIST were consistently higher than those of ANCILLES.

Correlations could not be computed for the c parameters since the true values were fixed at .25.

Correlations between ability estimates and true abilities were nearly equivalent for the two procedures. Increases were noted with increasing calibration test length but increases in sample size made trivial differences.

The mean-difference criteria suggested that both item parameters and ability estimates were biased somewhat. In general, ANCILLES produced more bias than LOGIST. Bias decreased with increasing test lengths and sample size.

Swaminathan and Gifford concluded that, although LOGIST produced slightly better estimation than did ANCILLES, it cost considerably more to run and the gain was probably not worth the cost. They further concluded that a and c parameters should not be estimated using tests containing 15 or fewer items.

Item Linking

Predicting, Equating, and Linking--A Clarification of Concepts

Scores from one test are often used to infer scores on a second test. Whether this inference is an act of predicting, equating, or linking will depend on the tests involved and the method used in making the inference.

Equating and predicting. Methods for equating test scores among different groups of people have long been available. Publishers of entrance examinations for educational institutions, faced with the need to change the examinations each time they were administered and aware that different types of people took the examinations in April and October, developed the means of assuring that a person of fixed ability would attain approximately the same score regardless of when the examination was administered. Formally, equating methods are procedures for expressing scores from two different tests measuring the same trait on a common score metric. The crucial requirement is that the tests measure the same trait.

Methods for predicting one test score from another have also long been available. The reason for giving entrance examinations in the first place was based on the empirical fact that scores on the entrance examinations predicted, to some degree, scores on classroom examinations. The difference between equating and prediction is that two tests do not have to measure the same trait to be candidates for prediction.

Statistical methods for equating and predicting come in both linear and non-linear forms. In the linear case, prediction is accomplished by linear regression. Equating is accomplished by a similar procedure in which a correlation of 1.0 between tests is assumed. Prediction uses the empirical data to estimate the relationship between the two traits. Equating assumes, not unreasonably, that a trait should correlate very highly (i.e., perfectly) with itself. The prediction equation is not invertible; a regression equation used for predicting test A from test B cannot simply be reversed and re-applied to predict test B from test A. The exception to this rule is when the correlation between tests is perfect. The assumption of perfect correlations made in equating allows the equating equation to be used for the inverse transformation.

If equating procedures are used for a prediction problem, the result will be less-than-optimal predictions. If regression is used for an equating problem, the result will be a lack of correspondence between test scores, which was the objective of equating in the first place.

Linking. Linking is a term which describes the act of equating at the item level. The objective in equating, as discussed above, was to put total test scores onto a common metric. Linking is used to describe the process of putting items from different tests on a common metric. Linking was first investigated as a means to an end of test equating (Fan, 1957; Swineford & Fan, 1957) and did not generate a great deal of research interest. More recently, as a result of adaptive testing applications, linking has become a legitimate end in itself. Adaptive testing item pools, because of their size, have had to be constructed by linking smaller sets of items together on a common parameter metric.

The objective of this project was to find efficient ways of linking test items. Much of the research available to date has been on equating rather than linking. There are close parallels between the two, however, and the following review will include equating as well as linking efforts. Prediction is a vast subject and will not be covered except to point out instances in which it was used appropriately as a linking or equating method.

Paradigms of Linking and Equating

Linking and equating paradigms can be categorized on two basic aspects: the design by which data are collected and the method by which the linking transformation is determined. Angoff (1971), in a classic survey of equating methodology, listed six major equating designs. In terms of data collection, these six designs can be grouped into two categories: designs assuming equivalent samples of examinees to achieve equation (Designs I and II) and designs employing an anchor test to achieve equation (Designs III, IV, V, and VI). Transformations, in Angoff's designs, are determined either through linear or curvilinear means. Marco (1977), in a recent survey, listed three data collection designs: (a) all items are given to a single group of examinees, (b) the same set of items is administered to different groups of people, and (c) an anchor set of items is common to all tests given to different groups of people.

There are, in fact, four basic data collection designs of potential utility for linking: (a) the equivalent-groups method, (b) the equivalent-tests method, (c) the anchor-group method, and (d) the anchor-test method. Angoff's first two designs are contained in the equivalent-groups method, and his latter four are examples of the anchor-test method. Marco's three designs are, respectively, a special case of the equivalent-groups method, a special case of the equivalent-tests method, and the anchor-test method.

In theory, IRT explicitly makes the relationship among item parameters, across groups, linear. There is thus no need to discuss the curvilinear transformation procedures. Reckase (1979) presented the most exhaustive array of linear procedures yet encountered. As will be discussed, however, only the one called the major axis procedure is an appropriate linking transformation method. Transformation methods thus do not offer much ground for research.

In theory, IRT item parameters are invariant, except for a linear transformation, across groups of individuals. The constants of the linear transformation necessary to change one metric to another (assuming a unidimensional pool of items), are simple functions of the means and standard deviations of the abilities of the groups under consideration. When items are calibrated, there are four values that are undetermined and must be arbitrarily imposed: the a and b parameter means and the ability mean and standard deviation. Among this group of four values, there are two degrees of freedom corresponding to unit and origin of the metric to be chosen. The unit can be specified by fixing either the mean a parameter or the standard deviation of the ability distribution. When one is fixed, the other is determined. The origin can be specified by fixing either the mean b parameter or the mean of the ability distribution. Again, when one is fixed, the other is determined. Any one of the values can be varied at will as long as the corresponding value is also appropriately adjusted.

As an example, assume that a set of items had been calibrated on a group of individuals and that the ability mean and standard deviation were set at zero and one, respectively. If desirable, the ability mean and standard deviation could be changed to 50 and 10. To do this, each ability estimate would be multiplied by 10 and 50 would be added. Also, the a and b parameters would have to be adjusted accordingly. In this case, the a parameters would have to be divided by 10 and the b parameters transformed by multiplying them by 10 and adding 50. The c parameter is evaluated at an infinitely low ability level and is thus not affected by the transformation (i.e., any finite linear transformation leaves negative infinity untouched). A linear transformation such as this could be used to set the mean and standard deviation of the ability distribution or the mean a and b values to any value without affecting the performance of the ICC model as long as both parameters were adjusted in the two pairs.

Item linking in IRT models consists of finding two common values (i.e., ability mean and standard deviation or item parameter means) in different sets of items given to different groups of people and then of determining a linear transformation that equates these values as well as the remaining two values which are determined by them. In the methods discussed in the next paragraphs, different sets of assumptions necessary to match values will be presented. The differences between the methods are in the groups chosen as the reference groups and in the parameters matched. The concept of the linear transformation to equate item parameters is the same for all methods.

Methods based on sampling. In the equivalent-groups method of item linking, a sample of examinees available for item calibration is randomly split into two or more groups, and each group is given a different set of items. It is assumed that the distributions of abilities are equal in the various groups; ability mean and standard deviation are the values matched across groups in this method. Parameters a, b, and c are estimated separately in each group, abilities are estimated, and ability levels and item parameters are simultaneously transformed such that the ability means and standard deviations of the groups are equal. The mean and standard deviation (i.e., origin and unit) of ability are arbitrary when items are calibrated and must be set to some values. Calibration programs (e.g., LOGIST or OGIVIA) typically set them to zero and one, respectively. In the equivalent groups method of linking, which assumes equal ability distributions, setting means and standard deviations equal, as is done by the program, puts all parameters on a common metric.

The equivalent tests method allows an item pool to be divided randomly into sets of items and these sets of items administered to different groups of examinees. It is assumed that the item subpools are equivalent, and thus the method derives from the concept of randomly parallel tests. Item parameter means are the values matched across groups, and no assumption is required about the distribution of abilities in the samples of examinees. As in the equivalent groups method, parameters $\underline{a}$, $\underline{b}$, and $\underline{c}$, as well as abilities, are estimated separately in each group. The difference is that the ability estimates and the $\underline{a}$ and $\underline{b}$ parameters are simultaneously adjusted such that the item parameter means, rather than the ability mean and standard deviation, are constant across groups (e.g., mean $\underline{a}$ of 1.0 and $\underline{b}$ of 0.0). Theoretically, the $\underline{c}$ parameter does not change across groups.

Methods based on anchoring. In the anchor-group method, a common group (i.e., anchor group) of individuals takes all items in the pool. Each subset of items is administered to a calibration group consisting of the anchor group and an additional group of examinees. The distribution of ability in the anchor group is taken as a standard, and no assumption of randomly sampled examinees or items is required. This method is conceptually very similar to the equivalent-groups method. Items are calibrated independently in each of the calibration groups as in the equivalent-groups method. The difference lies in the group of examinees on which the origin and unit of ability are established. In the equivalent-groups method, the mean and standard deviation of ability are assumed constant across calibration groups so the mean and standard deviation of ability in each of the groups is set to the same value. In the anchor-groups method, only ability in the anchor group is constant across calibration groups so, within each calibration group, a linear transformation of the item parameters is found which makes the ability estimate means and standard deviations within the anchor groups constant across calibration groups (e.g., 0.0 and 1.0).

The anchor-test method is based on a common set of items administered to all examinees. The anchor items are taken as the standard against which all other sets of items are calibrated. Parameters of the anchor test items are first estimated on the entire sample from the population of examinees. The mean and standard deviation of ability in this sample can arbitrarily be set to zero and one, respectively. Then for each subset of non-anchor test items given to a subgroup of examinees from the available population, item parameters and abilities are estimated. Each examinee in a subgroup will have an ability estimate from the anchor test items and another ability estimate from the non-anchor test items. Since the metric of the anchor test items is the standard, a transformation of item parameters of the non-anchor test items must be found which will make ability estimate means and standard deviations equal for both anchor and non-anchor test items. As was the case with the anchor-group method, no assumptions regarding the distribution of item parameters or abilities are required.

Composite network methods. The term network linking will be used to refer to any linking paradigm in which one of the anchor methods discussed above is used to simultaneously link items from more than two tests. Included in this category are the cascading schemes discussed by Angoff (1971) as well as the more complex networks described by Wright (1977) and Forster and Ingebo (1979). Conceptually, network procedures accomplish the same thing as the simple methods discussed above. They also provide advantages not available in the simple methods, however. Cascading schemes allow more efficient use of subjects when abilities are spread over a wide range. The more complex networks allow this and additionally allow independent checks on the links and evaluation of linking adequacy.

Criteria of Linking Adequacy

Item linking and item calibration are two psychometric activities that are intimately interrelated in practice. They are conceptually, however, two distinct operations, and it is important to recognize this fact when evaluating criteria for the adequacy with which each is done. Adequacy of calibration is evaluated by determining the accuracy with which the parameters of the items are estimated. The essence of IRT linking, however, is embodied in the linear transformation used to put items onto a common metric. This transformation is specified by two parameters: unit and origin. It is thus the accuracy with which these two parameters are estimated that determines the adequacy of the link. Estimates of the two parameters are subject to the same estimation quality criteria discussed above in reference to the item parameters: unbiasedness, efficiency, sufficiency, and consistency.

Few of the studies discussed below have given adequate thought to the criteria of linking effectiveness. In most cases, linking and calibration effects have been hopelessly confounded. In some studies of linking, no criteria that adequately reflect linking adequacy have been included. These deficiencies will be pointed out as the studies are discussed. More appropriate criteria will be presented later in this report.

Evaluation of Linking Techniques

Rasch Model. Of all the IRT models, the Rasch model is by far the simplest. It is a special case of the three-parameter logistic model which specifies that items can differ only in terms of difficulty. Graphically, this means that each ICC has the same slope but a different position to the right or left on the theta continuum. Although not a model of prime interest to the current research, because it fails to consider that guessing is possible in multiple-choice tests, most of the recent studies of linking and equating have been done using the Rasch procedure. A representative sample of these studies is thus reviewed below.

As in other logistic models, the Rasch ability parameters and item difficulty parameters (the only parameters in the Rasch model) are expressed on a common scale. Lack of an item discrimination parameter puts an additional restriction on the model in calibration: all items must be equally discriminating. In typical formulations of the model, the effective value of the common $\underline{a}$ parameters is 1/1.7 or about .59. If the actual value (in the logistic model frame of reference) is .59, the ability distribution will have a variance of 1.0. If the actual value is anything else, the variance will be other than 1.0. Similarly, if the average person ability is equal to the average item difficulty, or item easiness in Rasch terminology, the mean of the ability distribution (in the logistic frame of reference) will be 0.0.

Linking, as is commonly done with the Rasch model, consists of determining an additive constant to adjust both item easiness and ability values to a scale having a common origin. This is typically done in one of two ways. The first method requires that a common group of examinees respond to the item sets to be equated. Since the ability of the sample of persons is the same in both item sets, any differences in average ability computed from the different item sets are due to differences between the item sets. The second method requires that two groups of examinees respond to two item sets which share a common subset of items. In this method, the model states that because the common core of items should have the same average item easiness in both sets, any observed difference is due to differences in ability levels of the two groups in which the two sets of items are calibrated. An adjustment making the item easiness equal in the core items can be applied to the non-core items to place them onto the common scale.

In order for linking to be possible in this simple form, the discriminating powers of the items must be constant not only within tests but also across tests. Otherwise, only the means of the tests would be equated and not the variances. Most of the studies involving the Rasch model make the assumption of equal item discriminations across tests.

Several recent studies have investigated the utility of the Rasch model for the equating/linking of the National Board Medical Examinations. Bell (1979) used an anchor test to equate a 225-item Physician's Assistants Examination given in 1978 with a similar version given in 1976 (referred to here as the reference test). The anchor test was a 46-item set that had been included in all Physician's Assistants Examinations given since the testing program was begun. Bell evaluated two procedures in terms of their ability to answer two questions:

1. Is the ability level of current examinees higher than the reference group on which the reference test was originally calibrated?
2. Are the items on the current test more difficult than those on the reference test?

The procedures Bell compared were the Rasch model and several variants of linear raw score equating. For the Rasch procedure, each examination was calibrated separately. This yielded easiness parameters for each item set and ability estimates for each examinee group. Using a shift constant computed from the 46-item anchor test, ability scores from the current test were shifted to the scale of the reference test. The linear raw-score equating procedure began by estimating the mean and variance for both tests from the performances of the current group and the reference group on their respective tests and the combined (current and reference) group on the common items. These estimates were then used in a linear equation to yield a raw-score conversion. This procedure was not specified in detail but reference was made to Angoff's (1971) equating procedure for groups not widely different in ability. Bell concluded that although each procedure was capable of answering the question about the ability level of the current examinee group, only the Rasch model answered the question about whether the difficulty of the current items had increased. No discusssion was given as to the fit of the data to the Rasch model so judgment of the accuracy of the equating cannot be made. Due to the brevity of the paper, no more detailed inferences can be drawn.

Kelly (1979) discussed a large Rasch linking study in which items from two forms of a 1,000-item examination were linked together onto a common scale. The tests used, licensing examinations for medical doctors, were each composed of seven subtests of approximately equal length, assessing areas as diverse in content as biochemistry and behavioral science. Kelly made the assumption that these subtests all measured knowledge of medical science and were unidimensional enough in total to allow Rasch calibration. Statistical tests of this assumption, not described in enough detail to evaluate, reportedly supported its tenability.

Kelly described two studies. In the first, the seven subtests of a reference form of the test were administered to approximately 8,500 second-year medical students. Items in this test were all put onto a common scale by shifting subtest difficulty by an amount necessary to make ability estimate means zero for each of the subtests. The implicit assumption of equal item discrimination among subtests was apparently not tested. A second form of the test, the current form to be linked to the reference form, was given to approximately 3,000 second-year medical students. There were an unspecified number of common items between corresponding subtests in the two test forms. The linkage between the forms was established by first calibrating items of each subtest in the current form in the current group and then setting mean difficulties of the common items within subtests equal across the two forms. Uncommon items in the current test were put onto the reference test metric by adjusting them using the constant used to adjust the common items in the corresponding subtest. This resulted, given the assumptions, in a pool of 2,000 items all linked onto a common scale.

In the second study that Kelly described, both the reference test and the current test were first calibrated separately as 1,000-item homogeneous tests. Linking was accomplished by finding the constant that adjusted the common items to have equal mean difficulties in the two examinee groups. This was done in the same manner used for the subtests earlier. The difference here was that the entire test was linked at one time. This study was primarily descriptive rather than evaluative and, as such, provided no information on comparisons of linking designs. It did, however, illustrate two different designs. In the first study, linking was accomplished using a degenerate case of the equivalent-groups method (in which the groups were identical) and the anchor-test method. The second study used the anchor-test method exclusively.

The major flaw in Kelly's study is that it was purely descriptive rather than evaluative. It would have been informative, for example, to have a comparison of the two equating procedures using the same data. It seems reasonable to assume that both procedures would yield nearly the same results, but an empirical validation would be more convincing.

In the third study, sponsored by the National Board of Medical Examiners, Hughes (1979) used data from six tests given to different groups of examinees and placed the tests onto a common scale. Each test was composed of either 10 or 11 sets of six multiple-choice questions for a specific physician-patient interaction. The common-item links were thus composed of sets of questions, an arrangement that probably violated the local independence assumption of IRT.

The procedure for linking the six tests consisted of a complex network of common-item links. An iterative procedure computed estimates of each test's average difficulty on a common scale and expected values of the shift constant for tests having no common-item link. Two indices were proposed to identify inconsistent triads and links: a triad index and a link index. No information was provided about the distribution of these indices. Thus, only relative statements about the quality of the linking networks could be made. Although no conclusions were stated, use of the links and triad indices as diagnostic tools in evaluating the quality of Rasch linking was suggested.

Rentz and Bashaw (1975, 1977) applied item analysis and scaling methods of the Rasch model to data from the equating phase of the Anchor Test Study (Loret, Seder, Bianchini, & Vale, 1974) in the development of the National Reference Scale (NRS) for reading. The NRS was developed from seven widely used standardized reading tests consisting of vocabulary and comprehension subtests. There were two forms of each test, a primary and an alternate form. All 14 tests were chosen to be appropriate for grades 4, 5, and 6.

Seven pairs of tests were studied at each of the three grade levels. Each examinee responded to two reading tests. Each pair of tests was administered, counterbalanced, to two separate samples within each grade level yielding a total of 42 samples per grade level. In addition, each test was paired with its alternate form, counterbalanced within each grade level, and administered to 14 additional samples.

All tests at a single grade level were placed onto a common scale. Within each grade level, test pairs were calibrated as a single long test. The average item easiness was computed for each single test and the differences in averages were then computed for the test pair. These average differences were organized into matrices such that the lower half of the matrix contained differences from one order of testing and the upper half of the matrix, from the second order of testing. Row and column means were averaged, reversing the signs of the row means (due to reversed orders of administration), to obtain the equating constant averaged over order of administration. Tests were then placed onto a common scale defined by the Sequential Tests of Educational Progress--Series II (STEP-II) which was administered to all grade levels.

Comparisons of equated raw scores (i.e., number correct with no correction for guessing) from the Anchor Test Study and the Rasch study were made across samples from each study that took the same tests in the same order. For each comparison, the first test administered was taken as the base test. Conditional mean-squared errors were then computed for each base test score. For the comparisons reported, the differences between the equipercentile and the Rasch-based equated scores ranged from 0 to 3 raw-score points and were deemed inconsequential.

Slinde and Linn (1978, 1979) presented a set of studies designed to evaluate the adequacy of the Rasch model for vertical equating (i.e., equating where tests differ widely in difficulty and examinees differ widely in ability). In their first study (Slinde & Linn, 1978) response data from 1,365 examinees on a 36-item mathematics achievement test were used. Two tests of differing difficulty were obtained by dividing the 36-item test into two 18-item tests on the basis of the p-values of the items obtained in the group of 1,365 examinees. The average p-values of the tests were .665 for the easy test and .362 for the difficult test. The examinees were then divided into low-, middle-, and high-ability groups on the basis of their scores on the easy test.

Rasch item parameters were calculated for the total set of 36 items in the low group, the high group, and the total group (the middle group was reserved for later use). Ability estimates were then calculated for each of these groups (low, high, and total) using parameters obtained from each group in a crossed design. Mean differences between ability estimates derived from the easy test and the difficult test were then computed and compared.

When the total group ability estimates were calculated using item parameters obtained from the total group, the difference between means obtained from the easy and difficult tests was trivial. Similarly, when the high group mean was calculated using item parameters obtained from the high group and when the low group mean was calculated using the item parameters obtained from the low group, the differences were trivial. When items calibrated in the high group were used to estimate abilities in the low group or the middle group and when items calibrated in the low group were used to estimate abilities in the high group or the middle group, substantial differences in ability estimate means were found. Slinde and Linn interpreted this to mean that Rasch parameters were not really invariant and that Rasch equating procedures were not particularly useful for the problem of vertical equating.

Gustafsson (1979) criticized this interpretation. He suspected that the differences between means was due to regression artifacts which were due to the fact that Slinde and Linn had estimated abilities and subgrouped people on the basis of only 18 of their 36 items. Individuals would not be expected to perform, in a relative sense, as extremely in either direction on the entire 36 items as they did on the easy 18; therefore, a difference between means would be expected. To support his hypothesis, Gustafsson performed a computer simulation modeled closely after the Slinde and Linn study with the notable exception that the assumed invariance properties of the Rasch model were built in. His simulation showed that the parameter estimates obtained in the different groups were different but that this was due to a regression artifact and not to a lack of invariance.

He suggested that Slinde and Linn reanalyze their data, subgrouping individuals on the basis of their total test scores.

Slinde and Linn (1979) improved upon this idea by obtaining data from 1,638 examinees on two different tests including a 60-item reading comprehension test. The first test was used to independently subgroup examinees. The 60-item test was then split, on the basis of item difficulty, into two 30-item tests and their original study was essentially replicated. Their findings were that the mean differences disappeared in comparisons of the middle with the high group. Whenever the low group was compared with another group, the differences persisted. This finding was attributed to the effects of guessing. No allowance is made by the Rasch model for the possibility that correct responses can be obtained through guessing. When multiple-choice items are used, as was the case here, guessing undoubtedly happens and probably tends to bias the results. Most likely this was a more pronounced effect for the low ability group where subjects knew the correct answer less often and had more "opportunity" to guess.

Together these studies suggest that linear equating works as expected using the Rasch model but that problems may result if the model is used in groups of sufficiently low ability that guessing occurs with any frequency. Unfortunately, most items used in objective tests can be answered correctly by guessing and may often be used in environments where guessing is likely to occur. The three-parameter logistic model extends the Rasch model to account for guessing and thus may be more generally useful.

Three-parameter logistic model. In the three-parameter logistic model, as in the simple Rasch model, a linear equation is used to link parameters on one test to those on another. The one difference in the three-parameter case is the explicit addition of a scaling parameter to adjust for changes in unit as well as origin.

Three studies of linking using the three-parameter logistic model were of direct relevance to the present effort. One, a study by Reckase (1979), was of interest for two reasons: first he presented four methods of determining the linking transformation, and second, he attempted to determine acceptable numbers of items to be included in anchor tests for adequate linking to be possible. The four techniques for item linking he presented were: (a) major axis, (b) least squares, (c) least squares with outliers deleted, and (d) maximum likelihood.

The major-axis technique got its name from the fact that the parameter transformation equation was derived from the equation for the major axis of the ellipse formed by the data points of a bivariate plot of parameters of items in the tests being linked. In simpler terms, it amounted to a linear regression of the current parameters onto the reference parameters assuming the correlation to be perfect. Adjustment was made for unit and origin but no actual regression was performed.

The least-squares procedure was a regression procedure where the correlation was determined empirically rather than assumed to be perfect. As discussed earlier, this is not a legitimate linking method but rather a method of prediction.

The least-squares-with-outliers-deleted procedure presented was the same as the least-squares procedure, but items with parameters further than two standard errors from the regression line were deleted. Like the other least-squares procedure, this was not a legitimate linking method.

The maximum-likelihood procedure described by Reckase was really a version of the major-axis method. The procedure, as described, made use of the capability of the program LOGIST to treat items as "not reached" and ignore them in estimation of ability. What LOGIST actually does can best be illustrated in the simple paradigm in which two tests, with some of their items common, are given to two groups. For examinees taking the first test, items unique to the second are coded "not reached." For examinees taking the second test, items unique to the first are treated as "not reached." LOGIST estimates abilities for all examinees using all items "reached." This means that each examinee is scored on those items contained in the test taken. Using these ability estimates, item parameters are then estimated. Before the estimation process, which is iterative, can proceed to another stage, the ability estimates are scaled to a mean of zero and a variance of one. To do this, all item parameters must be appropriately adjusted. The adjustment is a major-axis transformation designed to make the parameters of the common items equal and the overall ability mean zero and variance one. Asymptotically, the same result should be achieved by an ordinary major-axis transformation following separate calibrations. For estimation, however, the maximum-likelihood procedure has the advantage of using all available data on the common items for each of the two separate calibrations.

Reckase used live-testing data obtained from administration of the Iowa Test of Educational Development (ITED) given to 1,000 Iowa school students from each of grades 9, 10, 11, and 12. The ITED consisted of seven subtests with a total length of 357 items. A principal-components analysis produced a sufficiently strong first component to suggest unidimensionality. The data were calibrated using each of three programs: (a) a Rasch model program written by Wright and Panchapakesan (1969), (b) LOGIST, a three-parameter logistic maximum-likelihood program (Wood, Wingersky, & Lord, 1976), and (c) ANCILLES, a three-parameter logistic minimum chi-square program (Urry, 1978).

This study was designed to evaluate the joint effects of linking method, calibration procedure, sample size, and anchor test size. As was discussed earlier, the major-axis method of determining a transformation was the only true equating method presented and discussion will be limited to that method. Sample sizes were 100, 300, 500, 1,000, and 2,000 obtained using a "systematic sampling procedure" from a total of 4,000 cases. Three levels of item overlap were chosen: 5, 15, and 25 items.

Four 50-item tests were linked in each condition. These tests were cascaded in the sense that, except for the first and last test, each test was linked to the previous test and the following test by two different sets of anchor items. Overlap among items in the two anchor sets in each test was permitted. Linking was performed sequentially: the second test was linked to the first, the third test was linked to the first two, and the fourth test was linked to the first three.

Each test was calibrated with each calibration program for each sample size, and each set of four tests was linked for each sample size and degree of overlap. Thus, for each linking there were 15 combinations of sample size and common item overlap. The reference against which linking adequacy was judged was a full calibration of the entire 357-item test using the full sample.

The adequacy of the linking was evaluated in three ways: (a) correlations between the linked parameter values and the total-test-calibration parameter values, (b) a sum-of-squared-deviations quality-of-linking index (Wright, 1977), and (c) scatterplots of linked parameter values versus total-test-calibration parameter values.

Results of the correlational analysis for the Rasch linking showed a predictable pattern of increasing correlations as sample size and number of overlapping items increased. No statistically significant changes in correlation occurred as the number of tests linked increased, but significance would have been difficult to judge because all correlations were near 1.0. The sum-of-squared-deviations quality-of-linking index was computed and reported for the Rasch model, but because the chi-square values (a transformation of this index) were significant, even when the correlations were of the order of .999, Reckase concluded that this index bore little relationship to the quality of linking. Therefore, this quality-of-linking index was not reported for the three-parameter models.

For the three-parameter calibration models, the correlations tended to follow the same increasing trend as sample size increased. No data were available for the 5- or 25-item overlap combinations; therefore, no conclusions could be drawn regarding trends with increasing item over-lap. From the correlational data reported, there seemed to be evidence to indicate that ANCILLES performed substantially better than LOGIST.

One problem is apparent in this study. Linking in an IRT model is an attempt to make a linear transformation of parameters from one metric to another. Correlations, the major criteria used in this study, are insensitive to differences between linear transformations. Although they provide information about the accuracy of calibration, they say virtually nothing about the adequacy of linking. The one criterion that is related to linking quality, squared error of estimate, was eliminated from consideration because it showed a difference where the correlations showed none.

As the data for the three-parameter model were not complete at the time the report was written, the effects of item-overlap could not be evaluated. Furthermore, as only one linking paradigm was presented (i.e., an anchor test design) no comparisons among methods were possible. Thus, the study served to clarify some issues regarding methods of transformation but did not provide any hard empirical data regarding linking design for the three-parameter model.

Ree and Jensen (1980), in a simulation study, investigated the joint effects of varying calibration group sample size and linking group sample size on the quality of the item parameter estimates. Simulating two tests with common items, a pool of 140 hypothetical items was specified. This pool was split into two tests of 80 items each. Twenty of the items were common to the two tests. The first test, T1, was taken as the reference test and the second test, T2, as the current test. Although not stated in the report, the program OGIVIA was used for calibration (Ree, 1980a).

Two groups of 2,000 hypothetical examinees each were generated from a standard normal population and a response vector for each examinee on one of the two tests was generated according to the three-parameter logistic model. Four samples of size 250, 500, 1,000, and 2,000 were drawn with replacement from each group and were used to calibrate the corresponding test. The major axis method of linking, described earlier, was then used to link parameters of the current test to the metric of the reference test.

Two criteria were considered in evaluating the quality of the parameter estimates. They were the correlations between true and estimated item parameters and the average absolute differences between true and estimated parameters. In the portion of the study explicitly discussing linking, only the average absolute differences were presented as correlations were expected to be misleading.

Both criteria behaved as might be expected from other research when accuracy of calibration was investigated separately in the two tests. Correlations for the a and b parameters increased and average absolute error decreased as sample size increased. No definite trend was obvious for the c parameter, however. It was estimated relatively poorly at all sample sizes but some improvement was noticeable as the sample size rose to 2,000.

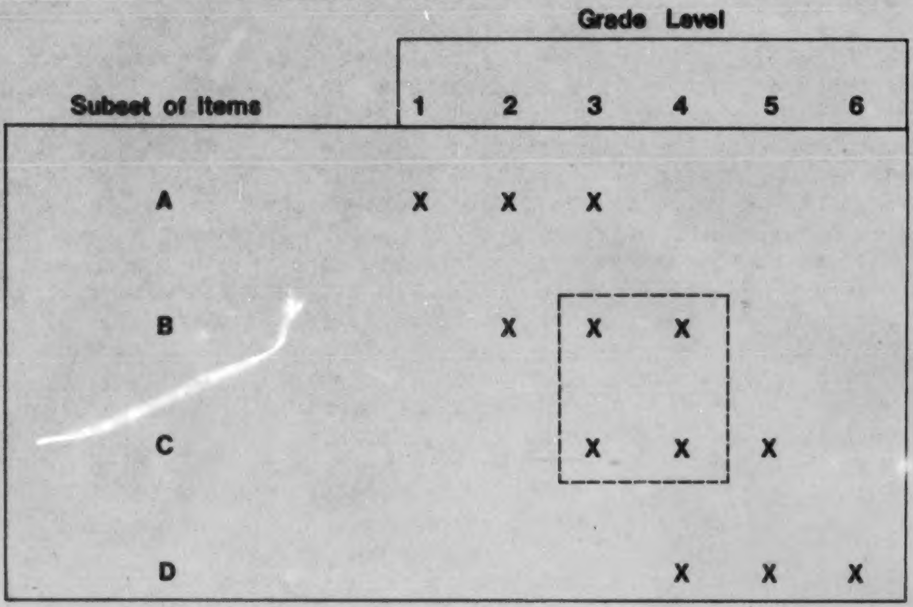
Linking adequacy was investigated at each of 16 combinations of reference and current group sample size for the a and b parameters. The c parameter, not in need of linking, was not considered. The expected trend toward decreasing error in the current test with increasing sample size was observed, for the most part, in the b parameters. As the size of the current test calibration sample increased, error in the b parameters decreased. There was a reversal with respect to the sample size used in calibrating the reference test: errors of estimation for the current-test b parameters were less for reference test calibration samples of 500 than for 1,000.

Errors in estimating a parameters did not follow such a reasonable pattern. Errors, as a function of reference test calibration group size, typically decreased with increasing size. Errors, as a function of current group size, were highest at a sample size of 250, lowest at a sample size of 500, and increasing from 500 to 1,000. It is this latter trend that was not expected.

An interesting comparison present in the data but not discussed was the relative quality of linking available from assuming equivalent groups of individuals when such an assumption is warranted (as it was in this study) compared to the quality of linking obtained from use of an anchor test. Since the calibration program assumed the ability metrics were the same for the two groups, the items were automatically linked upon calibration. Errors incurred in this linking were presented in the last column of Ree and Jensen's Table 5. When these results are compared to those obtained using the anchor test presented in their Table 6, it can be seen that the anchor test method was superior in only three of 16 sample size combinations for the a parameters and never superior for the b parameters. Thus, it appears, an explicit attempt to link items is not always necessary or desirable.

The third study of consequence to the present effort was a unique application of the three-parameter latent trait model by Sympson (1979). The procedure for placing items onto a common scale was unique in that it required neither overlapping groups of examinees nor overlapping sets of items. The data collection plan is schematically shown in Figure 2. Items were rank ordered in terms of difficulty and subtests were formed ranging from easy to difficult. Each subtest was administered to examinees at the grade level for which it was targeted and at the grade levels one level above and one level below that. Subtests were calibrated using responses of the three groups who took each subtest.

Figure 2. Sympson's Data Collection Plan



Subset of Items	Grade Level 1	2	3	4	5	6
A	X	X	X			
B		X	X	X		
C			X	X	X	
D				X	X	X

In order to place each subtest onto a common scale when there are no common items or common persons, Sympson suggested that if groups are randomly sampled from their respective populations, an equivalent-groups condition exists. This is indicated by the dashed box in Figure 2. The assumption of random sampling from a specified population implies, for example, that the group formed by combining individuals from levels 3 and 4 who took subset B was a random sample from the same "composite" population as the group formed by combining individuals from levels 3 and 4 who took subset C. Each pair of groups sampled from a common composite population was assumed to have the same mean and standard deviation on the underlying ability metric and thus comprised equivalent groups.

The paper was simply descriptive of the method and presented no data suggesting how well it worked. Reference was made to an unpublished simulation which apparently yielded favorable results. The paper's primary contribution to the current research is in its suggestion of a rather creative composite of simple procedures.

Conclusions

The research reviewed has been useful in suggesting potential methods of performing the act of item linking. Several data collection designs were suggested. Several methods of establishing the transformations were also suggested and served to clarify the fact that, for IRT models, only the major-axis procedure is appropriate. Finally, the studies reviewed suggested several criteria of linking adequacy. They served primarily to suggest a distinction between criteria of calibration and of linking adequacy and to suggest some candidates for linking-quality criteria.

The studies to date have not, singly or collectively, adequately dealt with the linking problem in general, however. Reckase (1979) attempted to compare methods of linking but his comparisons were primarily between transformation techniques not appropriate for linking. Ree & Jensen (1980) provided data relevant to the comparison of two data collection designs but the study was too small in scope to furnish much information regarding the linking problem in general. The remainder of the studies reviewed were primarily reports of how linking or equating had been accomplished for an applied problem and provided little insight into the general linking problem. The need for a broad investigation into the general linking problem seems obvious if linking is to be done accurately and efficiently.

The preceding discussion on the need to evaluate calibration and linking effectiveness separately was not intended to mean that calibration and linking are independent activities. The accuracy with which items are calibrated will have a definite effect on the accuracy with which items are linked. If, due to poor calibration, the ability levels of the groups are not accurately assessed, the transformation linking two groups will be in error. Similarly, the accuracy with which items are calibrated is, to some extent, dependent on the linking paradigm used.

It is thus important in a study of linking effectiveness to evaluate not only the adequacy of the link but also the adequacy of item calibration under the various paradigms. Ultimately, it is the accuracy with which the common-metric item parameters are estimated that will determine the quality of the tests resulting from these items, and this accuracy should be evaluated. Causes of inaccuracy in these parameters must, however, be evaluated by partitioning them into the effects due to calibration and the effects due to linking.

II. BASIC RESEARCH DESIGN

There are three general approaches to evaluating competing statistical or psychometric methods such as those considered by this project: a theoretical study, a real-data study, and a Monte-Carlo computer simulation (Weiss & Betz, 1973). In a theoretical study, a statistician or psychometrician, working from a basic statistical model, analytically derives the relevant characteristics of the various methods and then compares them. An example of this method was given by Lord (1971) in which he analytically derived several psychometric characteristics of a testing strategy. The theoretical method provides exact answers to theoretical questions but is usually limited to simple comparisons and comparisons made simple by restrictive assumptions.

Real-data studies answer different kinds of questions than do theoretical studies. Rather than answering questions about psychometric comparisons, they answer questions regarding characteristics of people and interactions of people with testing methods. They, in themselves, cannot answer questions such as which method best recovers true parameters because, in real data, the true parameters are never known. They are, nevertheless, essential in determining characteristics to use in theoretical or simulation studies and as a verification of the results of such studies.

A computer simulation is a modified theoretical study in which theory and data come together in a stochastic model simulating the responses of human examinees. Examples of a simulation study comparing testing methods are provided by Vale and Weiss (1975, 1978). Examples of simulation studies comparing calibration techniques are provided by Ree (1978, 1979). The simulation method is often preferred to real-data studies because true parameter values are known and more information can be collected more quickly. It is often preferred over a theoretical study because less restrictive assumptions are required. The simulation method is only as good as the theory underlying it and the reality of the parameters behind it, however.

To assure that the simulation results are meaningful, a simulation model must do two things: first, it must demonstrate a direct connection to the real-world problem that it simulates, and second, it must provide explicit answers to the questions of interest regarding the problem. The simulation models used in this project were anchored to the real world in two areas. First, the test items simulated were defined to be similar (in terms of their item parameters) to Armed Services aptitude items likely to be encountered in an actual linking problem. Second, the populations of individuals taking the tests were defined to be similar in ability to populations likely to take Armed Services tests. These procedures are described in the first of two sections below.

To address the research questions of interest adequately, the simulations and subsequent analyses must be properly designed and executed. In the second major section below, the research questions and the criteria used to evaluate the procedures are integrated into a concrete design for implementation of the study.

Development of Simulation Models

Specification of Items

Analyses of ASVAB item parameters. Two distinct sets of item parameter data were available for evaluation in preparation for the computer simulations. The first of these was an OGIVIA-produced IRT parameter set obtained from the subtests of an experimental version of Armed Services Vocational Aptitude Battery (ASVAB) Form 8 administered to Armed Forces Examining and Entrance Station (AFEES) examinees; a sample of 500 examinees was used to obtain the IRT parameters. Experimental Form 8 was a form of the ASVAB developed to parallel then-operational Form 7 (see Fruchter & Ree, 1977). The second set of data included the classical item parameters (i.e., the item-total score correlations and proportion correct) obtained from new Forms 8, 9, and 10 of the ASVAB administered, in a previous project, to groups of high school juniors and seniors. Each form was given to approximately 500 examinees. These parameters were transformed to IRT *a* and *b* parameters using Urry's method of simple approximation (Jensema, 1976). Because all items were four-alternative multiple-choice items, the *c* parameters were all set to .25

New ASVAB Forms 8, 9, and 10 differed from the old Forms 5, 6, and 7 (and, hence, from Experimental Form 8 discussed above) in that three of the original 12 subtests were eliminated, two subtests were combined, and two new subtests were added. Thus, there remained seven subtests in common between the two sets of available data. One of these subtests, Numerical Operations, was a speeded test and was therefore eliminated from consideration here because the logistic model is inappropriate for speeded tests. The six remaining subtests were Word Knowledge (WK), Arithmetic Reasoning (AR), Mathematics Knowledge (MK), Electronics Information (EI), Mechanical Comprehension (MC), and General Science (GS). In the new Forms 8 to 10, the lengths of five of these subtests were increased by 5 or 10 items; only the electronics test was shortened (by 10 items). See Table 1 for the numbers of items available in each of these subtests. These six subtests formed the basis for comparisons between Experimental Form 8 and the new Forms 8 to 10.

Table 2 presents summary statistics of items from the tests analyzed. The first four columns present values obtained for the first four central moments on the subtests of Experimental Form 8. The remaining four columns show values of the four moments obtained by pooling items from the new ASVAB Forms 8, 9, and 10.

Table 1. Number of Items in the Two Sets of Item Parameter Data

Subset	Experimental Form 8	New Forms 8, 9, 10 Within One Form	 Across All Forms Available
Word Knowledge (WK)	30	35	175
Arithmetic Reasoning (AR)	20	30	180
Math Knowledge (MK)	20	25	75
Electronics Information (EI)	30	20	60
Mechanical Comprehension (MC)	20	25	75
General Science (GS)	20	25	75

Note: For WK and AR, a total of 6 different forms existed for each subtest (e.g., Forms 8A, 8B, 9A, 9B, 10A, 10B); only the first five forms for WK were available for analysis and comparison. Only three distinct forms of each subtest existed for the last four subtests listed.

Mean proportions correct were higher on the new forms than on the experimental form. Values for each of the subtests clustered relatively close to the median values, however. The standard deviations were approximately equivalent across forms, again clustering near their medians. Comparing median skews, the proportions correct appeared to be nearly symmetric in both data sets. A relatively wide range of individual values was observed, however. Kurtosis was quite constant both within and across data sets; all proportion-correct distributions were quite platykurtic.

Biserial item-total correlations had relatively consistent means and standard deviations. There was some variation in skew within data sets. In the experimental form, values of skew ranged from -.872 to .012. In the new forms, the subtest skew ranged from -.432 to .089. Both medians were negative and not very different from each other. Kurtosis showed a wide range in the new forms, ranging from -1.009 to .390. It was less variable in the experimental form, ranging from -.822 to .120. The medians for the two data sets were not substantially different.

It was the IRT parameters, a, b, and c, that were most relevant to this project, however, as they were to form the basis for the simulation models. Mean a parameters were consistent within and across

Table 2. Item Parameter Summary Statistics from Experimental Form 8 and New Forms 8, 9, 10

		Experimental Form 8				New Forms 8, 9, 10 (Pooled)			
	Test	Mean	SD	Skew	Kurtosis	Mean	SD	Skew	Kurtosis
Prop. Corr.	WK	.602	.152	-.309	-.994	.716	.150	-.338	-1.024
	AR	.555	.168	.369	-.568	.656	.130	.121	-.766
	MK	.518	.141	.172	-.911	.619	.126	.111	-.664
	EI	.598	.126	-.455	-.750	.640	.160	-.230	-.955
	MC	.492	.165	.650	-.545	.625	.133	-.018	-.693
	GS	.511	.132	.178	-.997	.660	.148	-.450	-1.004
	Mdn	.536	.146	.175	-.830	.648	.140	-.124	-.860
Bis.	WK	.700	.113	-.717	.021	.670	.139	-.362	-.466
	AR	.667	.071	-.080	-.470	.646	.105	-.432	.390
	MK	.588	.124	-.744	-.608	.666	.084	-.089	-.862
	EI	.694	.089	-.872	-.145	.508	.136	-.097	-1.009
	MC	.625	.081	.012	-.822	.518	.110	.089	-.325
	GS	.629	.090	-.019	.120	.565	.112	-.044	-.891
	Mdn	.648	.090	-.398	-.308	.606	.111	-.093	-.664
a	WK	1.769	.536	-.124	-.180	2.171	.996	-.214	-1.621
	AR	1.816	.573	.789	.741	1.999	.904	.212	-1.498
	MK	1.602	.449	.706	.500	2.146	.848	.058	-1.581
	EI	1.486	.409	.444	-.190	1.183	.748	1.356	1.040
	MC	1.613	.388	-.129	-.713	1.116	.584	1.884	4.075
	GS	1.478	.627	1.019	1.433	1.439	.824	1.112	.012
	Mdn	1.608	.492	.575	.160	1.719	.836	.662	-.743
b	WK	-.005	.686	.312	-.810	-.333	.707	.309	-.375
	AR	.198	.772	-.484	-.572	-.126	.627	-.594	1.052
	MK	.510	.976	.525	-.016	.019	.545	-.226	-.063
	EI	-.014	.567	.098	-.886	.080	.908	.639	-.671
	MC	.577	.859	-.495	-.633	.070	.788	.219	.128
	GS	.413	.650	.456	-.027	-.079	.764	.825	-.376
	Mdn	.306	.729	.205	-.602	-.030	.736	.304	-.219

Table 2 (Continued). Item Parameter Summary Statistics from Experimental Form 8 and New Forms 8, 9, 10

		Experimental Form 8				New Forms 8, 9, 10 (Pooled)			
	Test	Mean	SD	Skew	Kurtosis	Mean	SD	Skew	Kurtosis
c	WK	.143	.067	.482	-.518				
	AR	.262	.114	.400	.635				
	MK	.293	.098	.754	-.411				
	EI	.170	.069	.626	-.467				
	MC	.287	.091	.938	.265				
	GS	.225	.113	-.368	-1.039				
	Mdn	.224	.094	.544	-.439				

Note: For the new Forms 8, 9, and 10, the c parameter was set to .25 for all items.

data sets; median values were 1.608 and 1.719. Standard deviations were quite variable within each data set, and the medians were markedly different (.492 vs. .836). The skews were typically positive but again somewhat variable. There were wide differences in kurtosis within and across data sets, as observed for the biserial correlation coefficients.

Part of the variability in the item statistics for the new ASVAB forms was undoubtedly due to difficulties with the item calibration procedure which caused a values to cluster at the upper limit. This clustering may be attributed to an artifact of the transformation procedure performed on the classical parameters from the new ASVAB forms. The theoretical relationship between the item-total biserial coefficients and the IRT a parameters is exponential, with high values for the former leading to very high values for the latter. At the upper end of the a distribution, then, the points are more spread out than they are at either the low end of the a distribution or the upper end of the distribution of biserials. (In this transformation procedure, the maximum a value was defined to be 3.20 and any transformed a which originally exceeded that value was set to 3.20. See Table 3 for the numbers of items which reached this maximum value.) This phenomenon would produce a distribution of a parameters which had a larger mean and standard deviation, was more positively skewed, and was somewhat more platykurtic than might otherwise be found. This, of course, is exactly what was observed for the new ASVAB forms.

The item parameters for Experimental Form 8, were produced by the OGIVIA program which relies on the same transformation for the initial parameter estimates. There are two crucial differences

Table 3. Numbers and Percentages of Items From the New Forms 8, 9, 10 With a Parameters Set Equal to the Maximum Value

Subtest	N in Subtest	N with Maximum a	Percentage with Maximum a
WK	175	72	41.14
AR	180	50	27.78
MK	75	21	28.00
EI	60	4	6.67
MC	75	3	4.00
GS	75	9	12.00
Total	640	159	24.84

between these parameters, however. The first is that the OGIVIA-produced a parameters from Experimental Form 8 were restricted so that the maximum a during the first and second stages was 2.40. During the ancillary corrections, however, there was no bound on the a parameters, and they were permitted to exceed 2.40 at this stage. The difference between the two procedures lies in OGIVIA's refinements of the item parameters based on values of the c parameters. For Experimental Form 8, as will be discussed below, the c parameters were quite variable. Although this was probably also the case with the "true" c's in the new ASVAB forms, all these c's were set to .25. The effects of these restrictions and of the c parameters on the estimation of a is reflected in the observation that the OGIVIA-produced a parameters did not cluster at the upper end of the distribution, and none were unreasonably large. Table 4 presents the numbers of items whose a parameters were equal to or exceeded 2.40 after the ancillary corrections; these relatively small values should be contrasted with the numbers of items with a parameters set to the maximum (3.20) in Table 3. For Experimental Form 8, only two items had a parameters exceeding 3.20.

The b-parameter means (Table 2) were slightly variable among subtests of the experimental form and quite constant in the new forms. Overall, the b parameters were slightly higher in the experimental form, indicating that either the items were more difficult or the AFEES examinees were less able than the high school students. Standard deviations were variable within data sets, but their overall medians were essentially equivalent. Skews ranged from -.495 to .525 in the experimental form and from -.594 to .825 in the new forms. Corresponding medians were .205 and .304. Kurtosis ranged from markedly flat to normal in the experimental form and from markedly flat to markedly peaked in the new forms; the kurtosis medians differed somewhat.

Table 4. Numbers and Percentages of Items From Experimental Form 8 With a Parameters Equal to or Exceeding 2.40

Subtest	N in Subtest	N with $a \geq 2.40$	Percentage with $a \geq 2.40$
WK	30	4	13.33
AR	20	3	15.00
MK	20	1	5.00
EI	30	0	0.00
MC	20	1	5.00
GS	19	2	10.53
Total	139	11	7.91

Note: One item from the original 20-item GS subtest was rejected by OGIVIA. Hence, IRT parameters were available for only 19 GS items.

Moments of the c parameters were calculated only for the experimental form as all c values were set to .25 in the new forms. Means and standard deviations were relatively consistent about their medians of .244 and .094, respectively. Skew was typically positive, with one exception. Kurtosis was variable, ranging from quite flat to somewhat peaked.

Table 5 presents intercorrelations among item parameters for Experimental Form 8 and new Forms 8, 9, and 10. For the new ASVAB forms where c was not estimated but, rather, set to .25, only the correlations between a and b could be calculated. The individual correlations exhibited considerable variation in all columns. The median of each column is presented at the bottom of Table 5. For Experimental Form 8, these medians were all essentially zero. For the new forms, the median a-b correlation was -.438.

Specification of a representative item domain. It appeared reasonable to assume that the item parameters summarized in Table 2 represented, with a few exceptions, a fair picture of the item domains likely to be encountered in the world of military testing. To form a basis for the simulations, a representative domain of items had to be specified. As with most scientific problems, there was a tradeoff between fidelity and practicality. The most faithful procedure would run all simulations on item sets representing each of the six subtests evaluated in Table 2. Practically, however, this would limit the number of simulations that could be run on any one item set. The approach taken in this project began by evaluating the item parameter data presented above to determine how far the six sets could reasonably be collapsed.

Table 5. Parameter Intercorrelations for
Experimental Form 8 and New Forms 8, 9, 10

Subtest	Experimental Form 8 a-b	 a-c	 b-c	New Forms 8, 9, 10 a-b
WK	.254	.311	.718	-.659
AR	-.152	-.154	-.607	-.173
MK	.027	-.334	.233	.037
EI	.300	.027	.315	-.625
MC	-.526	.011	-.494	-.349
GS	-.321	.026	-.104	-.527
Median	-.063	.018	.064	-.438

Note: The *c* parameter was set to .25 for all items in the New Forms 8, 9, 10. Therefore, only the correlation between the *a* and *b* parameters could be calculated.

The *a* parameters of the new forms were plagued by extreme estimates in nearly one-fourth of the items (see Table 3). Comparison of the first three tests with the last three tests hints at the extent of this problem. The safest route appeared to be to disregard the *a* parameters from the new forms and concentrate on those from the experimental form. A single domain with mean *a* of 1.6 and a standard deviation of .49 seemed reasonable. Skew and kurtosis values appeared to be nearly rectangularly distributed with few clusters. This suggested either one or six separate distributions. Six distributions seemed to be an extreme number to simulate just to capture differences in skewness and kurtosis. Median values were thus used. For the computer simulations, then, *a* was specified as having a mean of 1.60, a standard deviation of .49, skew of .58, and kurtosis of .16.

Although the medians of most of the *b* parameter moments were similar across the two forms, none of the distributions were appropriate for an adaptive testing item pool. Since adaptive testing is one of the major reasons for interest in IRT, the difficulty distributions were extensively altered for simulation. An item pool often considered ideal for adaptive testing has *b* parameters rectangularly distributed between *b*=-3.0 and *b*=3.0. Such a distribution has a mean of 0.0, a standard deviation of 1.73, a skew of 0.0, and kurtosis of -1.2. It is not unreasonable to expect item writers to be able to produce items similarly distributed. To allow for the practical consideration that more weight will undoubtedly be given to the center of the distribution, these specifications were relaxed somewhat. Thus, the *b* distribution used for the simulation was specified to have a mean of 0.0, a standard deviation of 1.5, a skew of 0.0, and a kurtosis of -1.0.

For input into the computer simulations, the c parameter distribution was specified to be as it was for Experimental Form 8. The parameters were: mean .24, standard deviation .09, skew .54, and kurtosis -.44. Because the median inter-parameter correlations were essentially zero for Experimental Form 8, uncorrelated parameter distributions were used for the simulations.

Item parameters were generated from the specified mean, variance, skew, and kurtosis using the power method described by Fleishman (1978). This procedure allows random numbers to be generated with the first four moments asymptotically specified.

Item parameters specified as described above did not always produce acceptable items. A few items were so extreme in difficulty that either all simulated examinees responded correctly or all responded incorrectly. When this happened, it was not possible to estimate parameter values for the item and it had to be discarded at the calibration phase. To prevent this from happening, items were rejected at an earlier phase when they were first generated if the expected proportion correct in a standard normal population was below .03 or above .97. This expected proportion correct was obtained from Equation 2 (From Owen, 1969, Eq. 6.2).

$$P = c + .5\ (1-c)\ [1-\mathrm{erf}(D)] \qquad [2]$$

where $D = b\ [2(a^{-2}+1)]^{-1/2}$

and $\mathrm{erf}(x) = 2\ (\pi)^{-1/2} \int_0^x \exp(-t^2)\ dt$

Rejection of items in this manner was expected to affect the distributions of the item parameters such that the moments would not be exactly as specified in the preceding paragraph. Since moments of the true parameters were needed for evaluation of some of the linking methods, a simulation was run to estimate these moments. In this simulation, 10,000 acceptable items were generated using the procedure described above. The first four moments were calculated for the three item-parameter distributions. For the a parameters, the mean, standard deviation, skew, and kurtosis, respectively, were 1.585, .488, .602, and .220; for the b parameters they were .227, 1.337, .079, and -.995; for the c parameters they were .240, .090, .527, and -.449. The only noticeable changes resulting from this rejection were in the b parameters; the mean rose slightly and the standard deviation and skew dropped slightly.

Specification of Ability Distributions

The objectives of the analysis of the AFEES ability distributions were threefold. The first was to obtain parameters of ability distributions for use in simulation models. Since one link between simulation and the real world is the ability distribution which generates the response vectors, the parameters describing this distribution should, as closely as possible, reflect the current AFEES examinee population. The second objective was to determine whether the AFEES examinees were sufficiently variable in mean ability to make item calibration more efficient by non-random assignment of experimental items. The final objective was to determine if the AFEES examinees were sufficiently similar that the equivalent-groups method could be effectively applied using the AFEES as the experimental sampling unit, even though that would violate a basic assumption of the method.

Examinee data available. The primary data available for analysis consisted of number-correct scores of 500 applicants from each of the 65 Continental United States (CONUS) AFEES on 12 subtests of ASVAB-7 randomly selected from tests administered during calendar year 1979. Six of the ASVAB-7 subtests were deleted from the analysis either because they were speeded tests or because they had been eliminated in the newer versions of the ASVAB. Fifty-six cases, in which keypunch errors were encountered, were deleted from the 32,500 cases available for analysis, leaving a total of 32,444 cases for further analysis. These deletions were essentially random and no single AFEES lost more than three cases to such errors.

Additionally, data from a sample of 500 applicants tested on an experimental version of ASVAB-8 were available in summary form. These data consisted of grouped frequency distributions of modal Bayesian latent trait estimates from the item calibration program, OGIVIA. They were collected during calendar year 1978.

Score data available. Ideally, latent trait estimates of ability should be used to evaluate the distributional characteristics of the underlying trait. The individual item response vectors needed to compute latent trait ability estimates were not available for analysis, however. The raw number-correct scores that comprised the primary data set were less than optimal for evaluation of ability distributions for several reasons. One major problem with using number-correct scores is that different response patterns can result in the same number-correct score. When test items differ in their characteristic functions, differing response patterns to a set of items, each containing the same number of correct reponses, can result in differing ability estimates. The effect of this is that the shape of the distribution of number-correct scores may differ from that of the underlying ability.

If IRT item parameters are available for a set of items, the test characteristic curve can be computed. This curve relates ability levels to true scores and can be used to approximate ability levels from number-correct scores. The item parameters were not available for ASVAB-7, however, and this transformation was not possible. The ability distributions were thus developed by simply standardizing the number-correct scores. The shape of the distribution of standardized scores would be correct if the test characteristic curve was linear. The degree to which this was true in the available data was not readily assessable.

The limited set of data available from the experimental form of ASVAB-8 did, however, provide an avenue for verification that the distribution shapes were reasonable. Although these data were not sufficient to draw any conclusions regarding differences among AFEES, they were adequate for evaluating the representativeness of the third and fourth moments.

Raw-score analysis. The parameters of the ability distributions for each subtest were estimated from the first four central moments of the total AFEES sample. The means and variances were set to zero and one, respectively, to facilitate subsequent analyses. Table 6 presents the skew and kurtosis for each ASVAB-7 subtest. With the exception of Word Knowledge and Electronics Information scores, which had slight negative skews, the remaining subtest scores had slight positive skews. Almost all subtest scores exhibited marked platykurtosis.

Table 6. Overall Skew and Kurtosis
ASVAB-7 Number-Correct Scores (N=32,444)

Subtest	Skew	Kurtosis
WK	-.114	-.991
AR	.162	-.850
MK	.328	-.717
EI	-.213	-.247
MC	.383	-.429
GS	.259	-.560
Median	.210	-.638

Because of the extreme flatness of the observed-score distributions, a check was made to ascertain whether this was due to outliers or whether it represented the true shape of the distribution.

The raw-score frequency distributions of a random sample of approximately 4% of the total AFEES sample for each ASVAB subtest are presented in Figures 3 to 8. It is apparent from the figures that the observed flatness was not an artifact caused by a clustering of scores at the endpoints. Thus the platykurtosis of the ability distributions is a realistic representation of the actual shape of the distribution. An earlier study by Fruchter and Ree (1977) describing the psychometric characteristics of experimental ASVAB Forms 8, 9, and 10 compared to operational Form 7B presented descriptive statistics from a sample of AFEES examinees similar to the present sample. Their results indicated the same trend toward platykurtosis as was found in this project.

Differences among AFEES. Two of the objectives of the AFEES evaluation centered on the determination of the differences in ability distributions among AFEES. Raw scores for all subtests were standardized by a linear transformation to a mean of zero and a standard deviation of one, as discussed above, to approximate the metric of a standard ability continuum. This standardization was done across all 32,444 examinees. The first four moments of these standard scores were then computed within each of the 65 AFEES groups.

Table 7 present summary statistics on the AFEES for each ASVAB subtest. The columns are the four central moments computed across AFEES (i.e., mean, standard deviation, skew, and kurtosis). The rows

Figure 3. Raw Score Frequency Distribution

Word Knowledge

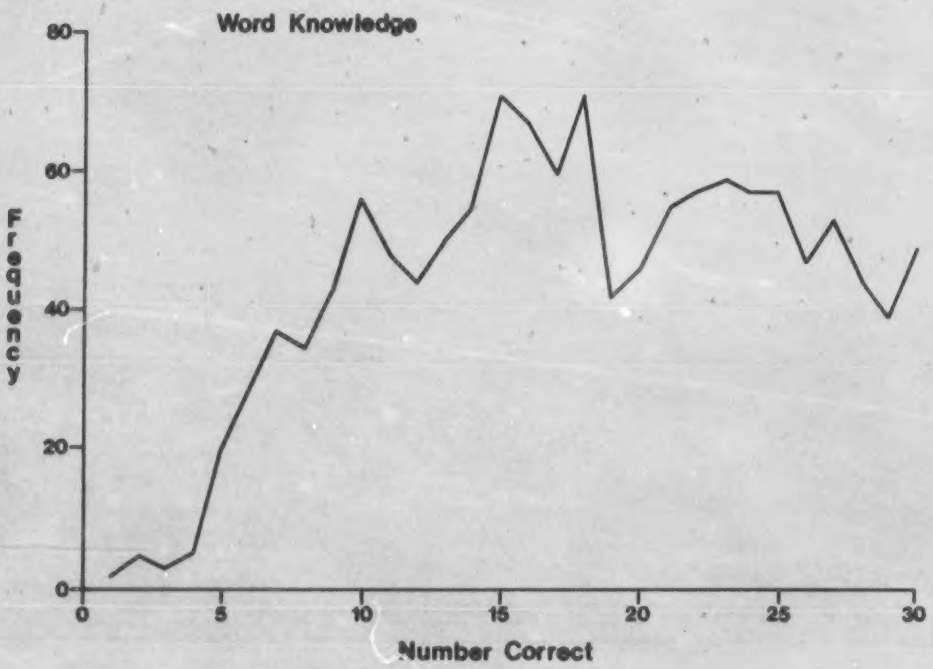


Figure 4. Raw Score Frequency Distribution Arithmetic Reasoning

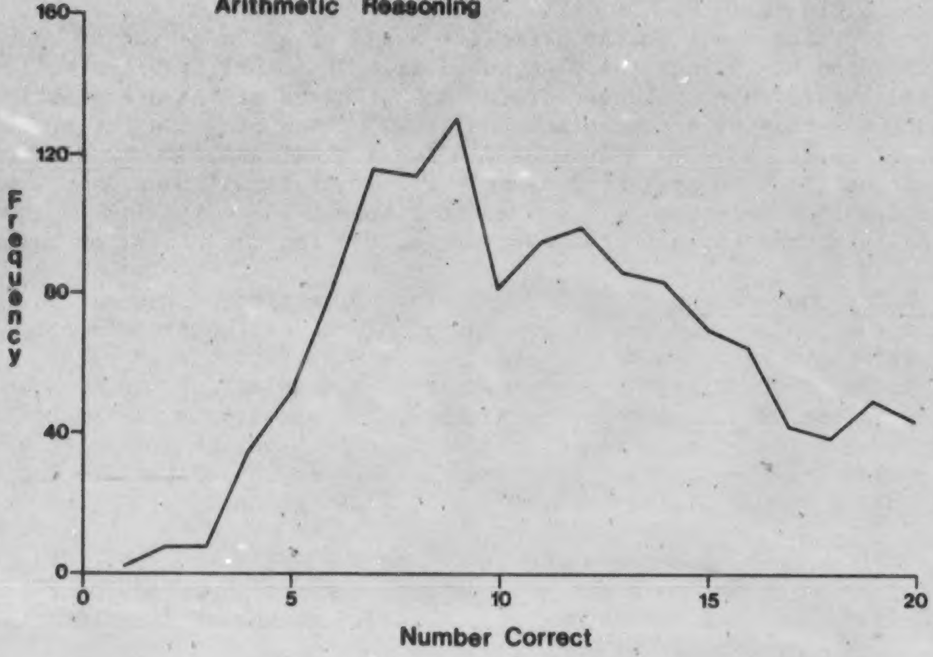


Figure 5. Raw Score Frequency Distribution Mathematics Knowledge

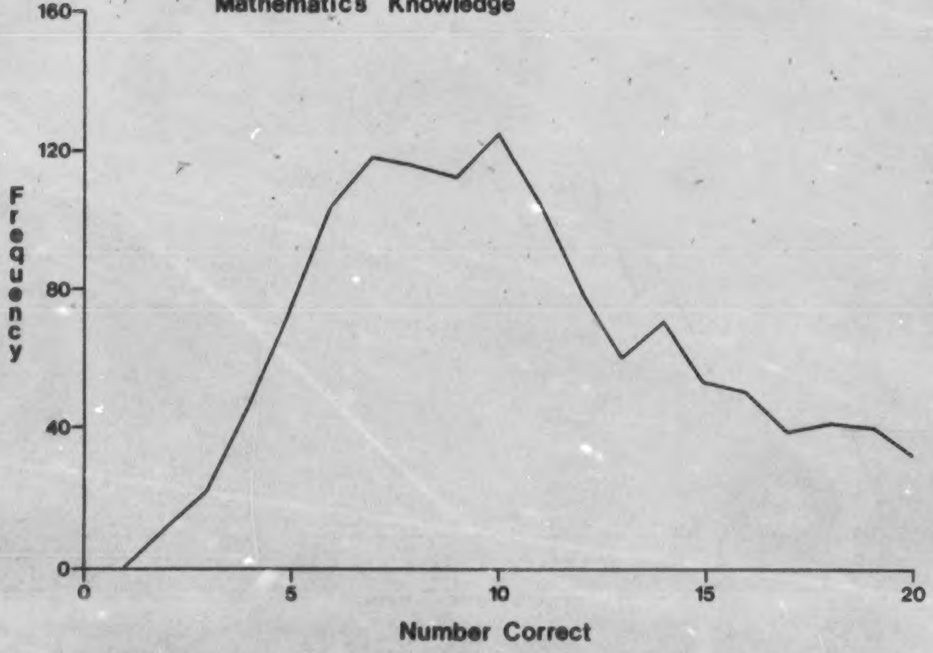


Figure 6. Raw Score Frequency Distribution
Electronics Information

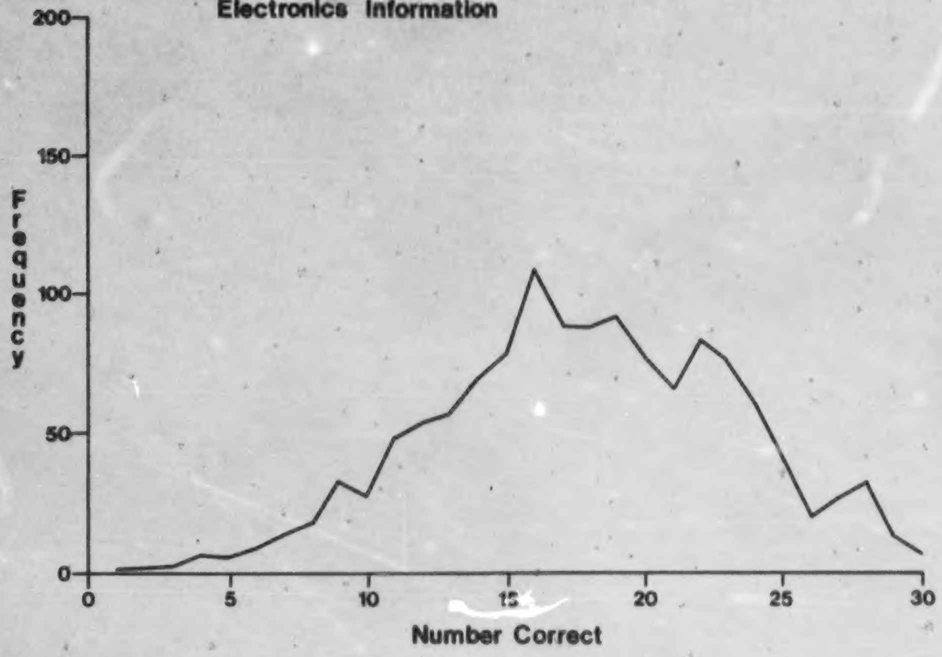


Figure 7. Raw Score Frequency Distribution
Mechanical Comprehension

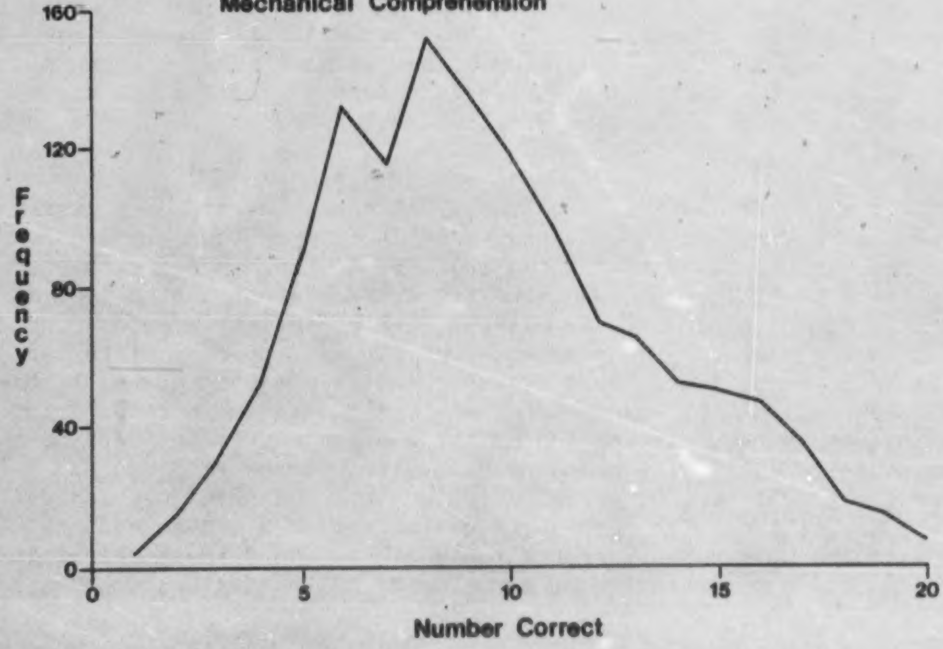
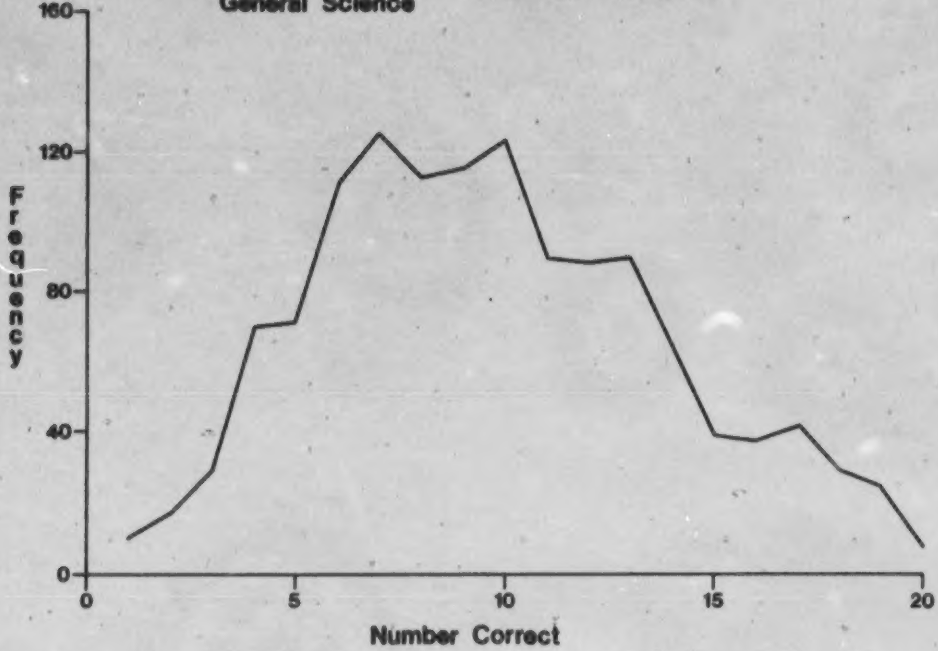


Figure 8. Raw Score Frequency Distribution
General Science



represent the ASVAB subtests and within each subtest, the mean, standard deviation, minimum and maximum of the first four moments. The mean of the means was zero in all cases since the computation was done on standard scores. The mean of the standard deviations was somewhat less than one. This is because part of the overall variance is due to variance among subgroup means which is not included in this calculation.

The standard deviations of the AFEES means and standard deviations are of interest in that they provide information regarding the error that will be introduced into the linked b and a parameters, respectively, if differences among the AFEES are not controlled in the linking process. If, for example, the equivalent-groups method was used and sampling was done non-randomly by assigning different booklets to each AFEES, these standard deviations are related to the root-mean-square (RMS) parameter error that would be introduced into the item parameters (the square of these values would be added to the mean-square error). The standard deviations of the AFEES means ranged from .201 to .244 which indicated that the AFEES were relatively homogeneous with respect to deviations about their central values. The mean-square error expected to be added to the linking error on the b parameters when sampling by AFEES was thus on the order of .040 to .060. Likewise, the range of linking error expected to be added to the *a* parameters was on the order of .001 to .003 (squared standard deviations of the AFEES standard deviations).

Table 7. Standard-Score Summary Statistics
Across AFEES for ASVAB-7 Subtests

Subtest		AFEES Moments by Subtests			
		Mean	SD	Skew	Kurtosis
WK	Mean	.000	.971	-.100	-.873
	SD	.235	.041	.245	.158
	Min	-.634	.876	-.512	-1.119
	Max	.385	1.060	.557	-.408
AR	Mean	.000	.975	.162	-.739
	SD	.222	.037	.219	.219
	Min	-.465	.852	-.350	-1.026
	Max	.428	1.056	.725	.157
MK	Mean	.000	.978	.321	-.620
	SD	.201	.049	.202	.306
	Min	-.340	.798	-.084	-1.078
	Max	.409	1.059	.718	.212
EI	Mean	.000	.972	-.188	-.193
	SD	.230	.049	.152	.253
	Min	-.544	.831	-.607	-.598
	Max	.384	1.056	.818	1.198
MC	Mean	.000	.969	.384	-.307
	SD	.244	.050	.196	.365
	Min	-.518	.794	-.073	-.833
	Max	.445	1.094	.820	.911
GS	Mean	.000	.974	.268	-.480
	SD	.225	.033	.167	.245
	Min	-.443	.882	-.097	-.867
	Max	.382	1.031	.680	.469

Comparisons of the overall skew and kurtosis given in Table 6 for each subtest with the skew and kurtosis for AFEES by subtest in Table 7 revealed virtually the same magnitudes and directions for the respective subtests. This indicated that the distributions of scores within AFEES were very similar in shape to the distributions over all

AFEES. Thus the four central moments computed for each subtest appeared to be reasonable estimates of the unknown true population values.

Modal Bayesian trait estimates. A parallel analysis was conducted on the available grouped frequency data provided by the IRT calibration program by computing the first four central moments for each ASVAB subtest. The formulas used to compute the moments were simply generalized versions of the formulas for ungrouped data where each element in the sum was the midpoint of its class interval weighted by the frequency of its occurrence.

As with the number-correct scores, the grouped modal Bayesian estimates exhibited consistent platykurtosis which ranged from -.607 for Arithmetic Reasoning to -.860 for Word Knowledge (see Table 8). Similarly, a slight skew was observed. Comparison of Table 8, which shows the four central moments for the ASVAB-8 modal Bayesian estimates, with Table 6 for the ASVAB-7 number-correct scores, indicates that the skews observed for the modal Bayesian estimates were similar to those of the number-correct scores observed over all AFEES. Agreement between data sets on observed kurtosis was also apparent. Both data sets agreed in direction and magnitude of the observed kurtosis.

Table 8. Mean, Standard Deviation, Skew, and Kurtosis of ASVAB-8 Modal Bayesian Ability Estimates (N=500)

Subtest	Mean	SD	Skew	Kurtosis
WK	.086	.854	.177	-.860
AR	.094	.805	.164	-.607
MK	.110	.736	.195	-.643
EI	.078	.807	.026	-.623
MC	.087	.785	.145	-.782
GS	.137	.729	.280	-.702

Overall, analysis of the modal Bayesian ability estimates tended to confirm the results of the number-correct score data and support the observation of flat ability distributions on ASVAB subtests. Although restricted to a fairly small sample (N=500) compared to the number-correct data, the modal Bayesian estimates were the preferred type of data. The results from these two rather disparate data sets tended to reveal the same general trends; therefore, the actual shapes of the underlying trait dimensions appeared to be adequately represented.

Specification of distributional parameters. To form a basis for the simulations, the ability data summarized in the preceding sections had to result in specification of a set of parameters to define the simulation models. To accommodate the simulations to be performed, two sets of ability parameters were needed. The first set required ability parameters for the overall AFEES distribution and the second set required ability parameters to describe each individual AFEES.

The data summarized consisted of six ASVAB subtests, representative of ability tests used by the Armed Services. To specify the parameters for the simulations, the first question to be answered was whether a single set of parameters could represent all of the tests or whether several sets would have to be included in the simulations. To answer this question, the skews and kurtoses of the overall distributions were of primary interest as the means and standard deviations were to be set to zero and one. Tables 6 and 8 allow comparisons between the skews and kurtoses of the ability distributions on the six subtests. Although many of the differences between subtests were statistically significant due to the large sample sizes, the absolute magnitude of the differences was relatively small. A general statement could be made that the ability distributions were, in most cases, symmetric and flat. The decision was thus made that a single subtest's ability distribution could be taken as representative of Armed Services ability tests.

The question remaining was how to choose the most representative test. Of two possible solutions, one was to use median values for the distributional parameters across the six subtests, while the other was to select a single test as representative and use its parameters throughout. It is possible, under the first approach, to get impossible combinations of parameters. Also, across AFEES, the parameters thus defined would have less variability than a typical set of parameters. A single test was thus chosen as representative of the ASVAB subtests.

To choose that subtest, the subtests were rank ordered according to their absolute deviations from the median of the overall skew and kurtosis values shown in Table 6. General Science and Arithmetic Reasoning ranked closest to the median for skew. General Science and Math Knowledge ranked closest to the median for kurtosis.

Across AFEES, it was essential that the test chosen as representative have representative variability in mean and standard deviation of the individual AFEES groups. The six subtests were thus rank-ordered on the standard deviation of their means across AFEES and the standard deviation of their standard deviations across AFEES. From the data in Table 7, it was determined that the typical tests in terms of variability of means were Electronics Information and General Science. In terms of standard deviations, the most typical were Math Knowledge and Word Knowledge.

Of the four comparisons, General Science was one of the most typical subtests in three out of four comparisons, the most of any subtest. Its parameters were thus selected for the simulation model. The overall ability parameters were thus mean of zero, standard deviation of one, skew of .259, and kurtosis of -.560. The four parameters from each of the 65 AFEES on the General Science test were used for individual AFEES simulations. These are listed in Appendix Table A-1.

Basic Data Sets

Four basic item linking paradigms were to be evaluated. It became apparent from review of the Armed Services calibration environment that practical administration constraints might, in a predictable fashion, violate a basic assumption of at least one of the paradigms. Specifically, the assignment of experimental test booklets to AFEES examinees would possibly be done non-randomly. In the limiting case, it is possible that each AFEES might receive a single form of a test booklet and, further, might be the only group to receive that booklet. Thus, two distribution schemes were simulated, the ideal case reflecting random distribution of test booklets and the worst case expected, that of non-random distribution.

The additional possibility existed that items might be calibrated on a selected group of examinees, such as those already in the Armed Services. A basic data set reflecting this situation was thus also developed.

Randomly sampled examinees. For the random-distribution case, a two-way grid composed of 12 combinations of test lengths of 20, 35, 50, and 65 items with examinee group sizes of 500, 1,000, and 2,000 formed the framework of the design. Within each cell, the specified number of examinees was randomly drawn from a standard ability population with a skew of .259 and a kurtosis of -.560. A sample of items was then drawn with parameters following the domain distribution specified in an earlier section. This process was repeated five times in each cell, with new random samples of examinees and items each time.

Systematically sampled examinees. The non-random procedure was similar to the random procedure except that for each replication, one of the 65 AFEES was randomly selected (with replacement) and its distributional statistics on the General Science test were used to describe the population from which examinees were drawn. In a real calibration design, the non-randomness of the sampling procedure would probably be less extreme. Each test booklet would probably be distributed over several AFEES groups. The exact distribution plan could not be predicted, however, and the limiting case was chosen to provide a bound to the errors that could be expected.

Selected examinees. One row of the basic matrix corresponding to 1,000 examinees was simulated at the standard test lengths of 20, 35, 50, and 65 items for the selected examinee condition. As with the other conditions, five replications were done in each cell. In this condition, however, 1,500 examinees were generated and sorted on the basis of the number-correct score. One thousand individuals with scores at or above the score of the individual ranked 1,000th were selected. This procedure was done to simulate examinees selected on the basis of a cutting score and the cutting score was chosen to be similar to that used by Ree (1979).

Composite sets of items. To evaluate the effects of linking procedures, items from more than one calibration must be combined and linked. To facilitate this evaluation, two types of composites were assembled from the basic data sets. In the homogeneous condition, the five sets in each cell of each 3x4 or 1x4 matrix were linked together. In cells containing 20-item sets, 100 items were linked together; in cells containing 65-item sets, 325 items were linked together. Composite sets so assembled provided data regarding linking adequacy when all sets included were homogeneous with regard to test length and size of calibration group.

The second type of composite, the heterogeneous condition, was formed by selecting 20 items from one set of each of the 12 cells of the 3x4 matrix to form a set of 240 items. Items beyond the first 20 in a set were ignored. This procedure resulted in five composites from each matrix, one corresponding to each replication within the cells. This type of composite yielded data regarding linking adequacy when sets included were heterogeneous with respect to test length and calibration group size.

Calibration of items. For each of the 140 administrations enumerated above, item responses were generated using true ability levels and true parameters according to the following algorithm:

1. The probability of a correct response to an item, given an individual's ability and the true item parameters, was calculated using Equation 1.

2. A random number from a rectangular distribution on the range from zero to one was drawn.

3. A response of "correct" was assigned if the probability exceeded the random number. Otherwise, a response of "incorrect" was assigned. (See Ree, 1980b, for a more detailed description of this type of procedure)

The item response data thus created were used as input to the item calibration program OGIVIA. This program provided item parameter estimates and modal Bayesian ability estimates (using a standard normal prior ability distribution).

For each of the administrations, the following statistics were recorded:

1. The first four moments of the population ability distribution.
2. The true parameters for each of the items.
3. The estimated parameters for each of the items.
4. The true ability level for each examinee.
5. The estimated ability level for each examinee.
6. The response of each examinee to each item.

These data formed the basic data sets used for analyses of the four basic linking methods. How the same data were used for the four different linking methods is described below.

Evaluative Criteria

Three categories of evaluative criteria were used to evaluate the adequacy of calibration and linking. The first category included the usual fidelity-of-estimation criteria used in previous studies. They were used in this study to provide simple indices of estimation accuracy and to provide a means of comparing the results of this study with those of previous studies.

A study of calibration and linking must consider that, ultimately, the interest will be in the effects of different techniques on the estimation of ability. Fidelity-of-estimation criteria do not afford any direct inference regarding accuracy of ability estimates. To ameliorate this problem, the last two categories of criteria evaluate the asymptotic (i.e., infinite test length) characteristics of ability estimates and the efficiencies with which various techniques approach these characteristics.

Fidelity of Parameter Estimation

Bias. Perhaps the most basic of the fidelity criteria is bias in the distributions of the item parameters. To assess the bias in the distributions of the parameters, means and standard deviations of the true and estimated parameters were calculated for all conditions of interest. The biased formula for the standard deviation was used, as it was throughout this research.

Absolute error. The mean absolute difference between true and estimated parameters was calculated and is referred to throughout this report as the absolute error. Algebraic error or bias may cancel out even though severe errors of estimation exist. Absolute error is one method used to eliminate this cancelling effect.

Root-mean-square error. Root-mean-square error is an index similar to absolute error except it is computed by taking the square root of the mean of the squared differences between true and estimated parameters. The primary difference in effect is that the root-mean-square index weights the extreme deviations more heavily than does the absolute index. Root-mean-square error was calculated for all conditions of interest.

Correlations. Correlations between true and estimated item parameters were calculated. The simple Pearson product-moment correlation was used. This index can be thought of as a complement to indices of algebraic bias. The bias indices are sensitive to changes in the location of the distribution of parameters. The correlation is sensitive to differences in relative position between corresponding true and estimated parameters.

Characteristics of Asymptotic Ability Estimates

Most of the desired knowledge that pertains to the ability to estimate a trait can be indexed by the bias and the precision with which the trait is estimated. In an effort to evaluate the bias due to calibration it is helpful to think of two trait metrics for the given trait of interest. The theta (θ) metric can be defined as the absolute or criterion metric on which the true parameters are anchored and along which the response probabilities are accurately described by the model incorporating the theta level and the item parameters. A second metric, gamma (Γ), can be described as a one-to-one transformation of the theta metric produced by scoring item responses using item parameters other than those true parameters of the theta metric. The gamma level corresponding to a given theta level could be determined, conceptually, from administering a test scored using the errant parameters an infinite number of times. Each theta value would thus asymptotically converge on a single gamma value. The difference between gamma and theta at any value of theta could be defined as the bias due to use of the errant parameters.

Practically, it is impossible to administer infinite-length tests or to repeat a finite-length test an infinite number of times. The theta-gamma transformation can be determined by more practical means, however. The maximum likelihood estimate of theta, which is asymptotically unbiased, can be obtained by finding the root in theta of the following equation given by Birnbaum (1968, p. 459):

$$\sum_{g=1}^{m} a_g \Psi[Da_g(\hat{\theta}-b_g)] - \sum_{g=1}^{m} \frac{w_g[\hat{\theta}]u_g}{D} = 0 \qquad [3]$$

where: $D = 1.7$

$$w_g[\hat{\theta}] = Da_g \Psi[(Da_g(\hat{\theta}-b_g) - \ln(c_g)] \qquad [4]$$

and $u_g = 1$ for a correct response to item _g_ and 0 otherwise.

If each item were repeated _r_ times, Equation 3 could be written as:

$$\sum_{g=1}^{m} \sum_{h=1}^{r} a_g \Psi[Da_g(\hat{\theta}-b_g)] - \sum_{g=1}^{m} \sum_{h=1}^{r} \frac{w_g[\hat{\theta}]u_{gh}}{D} = 0 \qquad [5]$$

or

$$r \sum_{g=1}^{r} a_g \psi[Da_g(\hat{\theta}-b_g)] - \sum_{g=1}^{m} \frac{w_g[\hat{\theta}]}{D} \sum_{h=1}^{r} u_{gh} = 0 \qquad [6]$$

or

$$\sum_{g=1}^{m} a_g \Psi[Da_g(\hat{\theta}-b_g)] - \sum_{g=1}^{m} \frac{w_g[\hat{\theta}]}{D} P_g = 0 \qquad [7]$$

where P_g = the observed proportion of correct responses to item _g_ in _r_ repetitions.

If the number of repetitions were allowed to become infinite and the three-parameter logistic model holds,

$$P_g = P_g(\theta) = c_g + (1-c_g) \quad [Da_g \, (\theta-b_g)] \qquad [8]$$

Computing P_g as above, the root of the likelihood equation is found at $\hat{\theta} = \theta$. If, however, P_g is calculated using θ and the errant item parameters $\underline{\hat{a}}_g$, $\underline{\hat{b}}_g$, and $\underline{\hat{c}}_g$, the root of Equation 7 is found at $\hat{\theta} = \Gamma$. If the errors of calibration are zero or the estimated parameters are consistent with the true parameters, the transformation of theta to gamma will be linear. When this is not the case, as in almost all real calibration situations, the transformation will be non-linear.

The function transforming theta to gamma completely describes the asymptotic effect of item parameter error on ability estimation. This empirical function has no simple descriptive parameters, however, and a method to condense many functions into table values was needed for this research. To accomplish this, a standard normal density function was taken as a reference theta population and the descriptive parameters of the corresponding gamma population were tabulated. Methods of calculation are described below.

Mean and standard deviation. For each calculation of the mean and standard deviation of gamma, 47 theta values equally spaced between -4.6 and 4.6 were chosen. At each of these values the standard normal density, the gamma value, and the squared gamma value were obtained. The gamma and squared gamma values were each numerically integrated jointly with the density using Simpson's one-third rule of quadrature to obtain the expected value of gamma and the expected value of gamma squared. The mean was taken as the former. The standard deviation was obtained by using the formula for expected values. To accommodate numerical limitations of the computer used, gamma was bounded between -5.0 and 5.0.

Absolute and root-mean-square error. Mean absolute and root-mean-square errors were calculated in a manner similar to the mean and standard deviation. At each of the 47 theta points, the absolute and squared differences between theta and gamma were calculated. The expected values of these quantities were obtained through joint numerical integration with the normal theta density function. The expected absolute error was the mean absolute error. The root-mean-square error was taken as the square root of the expected value of the squared difference between gamma and theta.

Correlation. The correlation between theta and gamma was computed as an index of linearity of the transformation. At each of the 47 theta values, the cross-product of theta and gamma was computed. Since all of the joint theta-gamma density falls along the regression function, this cross-product, jointly integrated with the normal theta density, produces the expected cross-product. The correlation between theta and gamma was computed from this value and the known and previously computed means and standard deviations of the theta and gamma distributions.

Efficiency of Ability Estimation

Although the transformation function provides a measure of the bias incurred through use of errant parameters, it tells little about the precision with which the parameters permit an estimate of the trait levels. An index closely related to precision of estimation is the statistical or Fisherian information. For a given test scoring function at a specified level of a trait, theta, this information can generally be expressed as the ratio of the squared derivative of the expected value of the scoring function to the variance of the scoring function at the specified level of theta:

$$I(\theta) = \frac{\left[\frac{d}{d\theta} E(x|\theta)\right]^2}{\sigma^2_{x|\theta}} \qquad [9]$$

When the score, x, is a linear combination of 0-1 item responses, the components of the information equation can be written as:

$$\frac{d}{d\theta} E(x|\theta) = \sum_{g=1}^{m} \frac{d}{d\theta} w_g E(u_g|\theta) \qquad [10]$$

$$= \sum_{g=1}^{m} \frac{d}{d\theta} w_g P_g(\theta)$$

$$= \sum_{g=1}^{m} w_g P'_g(\theta)$$

where

$$\sigma^2_{x|\theta} = \sum_{g=1}^{m} w_g^2 P_g(\theta) Q_g(\theta) \qquad [11]$$

w_g = a weight assigned to item g

and $P_g'(\theta) = (1-c_g)\, Da_g \psi[Da_g(\theta-b_g)]$.

Birnbaum (1968) discussed choosing the weights to be best or "locally" best in the sense that they would make the information of the linear combination maximal at a given value of theta. In cases where guessing is not possible, these weights are simply:

$$w_g = Da_g \tag{12}$$

In cases where guessing is effective, the weights change as a function of theta and are given by Equation 4 above. Weights obtained for a given level of theta would, when used in linear combination, provide maximum information for making discriminations between two theta levels arbitrarily close to the theta level of interest. When true item parameters are used, information computed in this manner is equal to the test information at the theta level of interest obtained by summing the item information values at that point.

The information in any linear combination can be evaluated; therefore, it makes sense to evaluate the information available at a given level of theta from items with errant parameters by evaluating the information in the linear combination obtained by using the locally best weights obtained through the errant parameters. This is done for a given theta level by first finding the corresponding gamma level. Weights are then determined using this gamma level in place of theta in Equation 4 and substituting the errant parameters for the true ones as in Equation 13:

$$Q_g(\hat{\Gamma}) = D\hat{a}_g \psi[D\hat{a}_g(\hat{\Gamma}-\beta_g) - (\ln \hat{c}_g)] \tag{13}$$

The information can then be determined by substituting $Q_g(\hat{\Gamma})$ for w_g in Equations 10 and 11. This information is interpretable on the same scale as the true information, and the relative information of tests using true and errant parameters can be obtained by taking their ratio. The reciprocal of this ratio can be interpreted as the relative numbers of items with true and errant parameters necessary to achieve an equivalent level of measurement precision at the specified trait level.

Information. The information function produced by the method described above is nearly as awkward to work with as the regression functions described earlier. The information function data were thus condensed in a similar manner. For each condition of interest, information was calculated at the 47 theta points. Expected information was then obtained by jointly integrating these information values with the standard normal density function. The resulting value represented the average amount of information that would be extracted by the test for an examinee selected at random from a standard normal population. To provide a basis for comparability, information per item is presented throughout this report.

Relative efficiency. When comparing information extracted by different procedures, the comparison is often done in terms of a ratio. The ratio of information from two tests is an index of relative efficiency. If the ratio of Test A information to Test B information is .80, Test A is 80% as efficient as Test B. Test B would achieve an efficiency equivalent to that of Test A with only 80% as many items as it currently has.

Whether an index will indicate calibration or linking error is dependent, in large part, on how it is applied. The indices presented thus far have all been discussed as indicators of calibration error. The underlying concepts and the indices themselves may, however, be used to evaluate linking errors by applying them to the case where multiple sets of items are calibrated separately and then linked together.

The effects of calibration and linking errors are difficult to separate using fidelity or asymptotic ability indices. They can be readily separated using the efficiency indices, however. Loss in efficiency is caused only by relative errors of calibration, not by constant errors. A linking error exists when the unit and origin of the trait resulting from the item parameters differ from the true unit and origin of the trait. Linking errors are constant within an item set; thus, they result in no loss of efficiency and are not usually considered a problem when all items are calibrated as a single set. If, however, two or more sets of items are calibrated separately and then combined into a single pool, errors constant within each set are now relative in the combined pool. The result will be a loss of efficiency.

Loss of efficiency in a single item set is due to calibration error. Loss of efficiency in a combined pool is due to both calibration and linking errors. The index of efficiency used in this study was information, and information is additive. If information contained in the combined pool is subtracted from the total information contained in the individual pools, the value remaining is the information lost as a result of linking. The ratio of the information available using the linked parameters to the information available using the true parameters yields an efficiency index of the linked items. The ratio of the information available from the linked parameters to the information available from the estimated parameters within sets yields an efficiency index of the linking procedure.

III. EVALUATION OF THE BASIC DATA SETS

Three basic data sets comprised the data on which most of the analyses reported here were based. Evaluation of these data served two purposes. First, they provided baseline data free of linking error for comparison in later phases of the study. Second, the data provided substantial information regarding the characteristics of the calibration procedure used (i.e., OGIVIA). These data allowed a more comprehensive analysis than was available from previous research because the evaluative criteria provided were both more extensive and more closely related to a test's capacity to estimate ability.

As will be the case with all analyses presented, each data set will be discussed separately. Within the discussion of each set, the three categories of evaluative criteria presented in the previous section will be discussed.

Randomly Sampled Examinees

Fidelity of Parameter Estimation

Table 9 presents parameter bias statistics for each of the three parameters, $\underline{a}$, $\underline{b}$, and $\underline{c}$, for the randomly sampled calibration groups. Bias, as used in this table, is the mean of the estimated parameters minus the mean of the true parameters. Means of values obtained from five calibrations are presented for each of the 12 cells in the center of each section of the table and row and column simple averages are presented in the margins.

As can be seen from the first section of the table, the $\underline{a}$ parameters exhibited substantial bias at short test lengths. At a length of 20 items, the estimates were high by approximately .6 units. This bias proceeded smoothly to zero by a test length of 65 items. No consistent change was observed in the amount of bias as the number of examinees in the calibration group increased from 500 to 2,000.

The $\underline{b}$ parameters exhibited relatively little bias in any of the 12 cells. The highest was .155 in the 20-item tests calibrated on 500 examinees. As shown by the marginal averages, bias decreased slightly with increasing test length and sample size. The decrease was very slight, however, and as can be observed from the individual cell entries, was by no means consistent. It may be observed that the errors for the $\underline{b}$ parameters were smaller than those for the $\underline{a}$ parameters. These comparisons are not readily interpretable, however, because the $\underline{a}$ and $\underline{b}$ parameters are on different scales.

Bias in the $\underline{c}$ parameters was also quite small. No obvious trend with respect to group size was observed but bias did appear to

Table 9. Item Parameter Bias
Basic Data Set--Randomly Sampled Examinees

	Sample	Test Length				
Parameter	Size	20	35	50	65	Average
a	500	.594	.292	.095	-.029	.238
	1000	.623	.232	.094	.009	.239
	2000	.581	.248	.079	.017	.231
	Average	.599	.257	.089	-.001	
b	500	.155	.121	.098	.102	.119
	1000	.114	.123	.129	.099	.117
	2000	.154	.089	.066	.071	.095
	Average	.141	.111	.098	.091	
c	500	.017	.024	.001	.006	.012
	1000	.014	.023	.011	-.003	.012
	2000	.033	.011	-.004	-.001	.010
	Average	.021	.020	.003	.001	

decrease with increasing test length. Although not as consistent as with the a parameters, this decrease was fairly consistent with increasing test length.

Table 10 presents correlations between true and estimated item parameters for the randomly selected calibration groups. Each cell entry represents Fisher's r-to-z average of correlations obtained independently in each of five calibrations. The marginal values are, likewise, r-to-z averages of the cell averages.

These correlations ranged from .435 to .684. Slight increases in correlations between true and estimated a parameters with increasing test length and calibration group size are apparent in the first section of Table 10. The increases were not markedly consistent, however, as may be observed both in the marginal and the cell entries.

Similar observations can be made regarding trends in the b-parameter correlations. Slight but consistent increases were observed in the marginal values. The individual rows and columns did not all exhibit the same consistency, however. Although the increases were slight (from .985 to .990), it should be noted that slight increases

Table 10. Parameter Correlations
Basic Data Set--Randomly Sampled Examinees

Parameter	Sample Size	Test Length 20	35	50	65	Average
a	500	.435	.505	.632	.647	.561
	1000	.645	.612	.673	.560	.624
	2000	.460	.643	.684	.659	.618
	Average	.520	.590	.664	.624	
b	500	.978	.984	.986	.988	.984
	1000	.989	.987	.989	.992	.989
	2000	.986	.992	.992	.990	.991
	Average	.985	.988	.989	.990	
c	500	.377	.460	.465	.432	.434
	1000	.431	.555	.560	.605	.541
	2000	.383	.555	.555	.529	.509
	Average	.397	.525	.528	.526	

are important in correlations as near to 1.0 as these. The correlational data presented here suggest that the b parameters are extremely well estimated at all combinations of test length and calibration group size considered.

Relatively consistent improvements in the c-parameter correlations were observed as test length increased up to a length of 50 items. At a length of 65 items, two of the three correlations dropped slightly. Improvement with increasing sample size increased to a size of 1000 examinees. Increasing the sample size to 2000 resulted in no improvements. Overall, the c-parameter correlations were slightly lower than those of the a parameters. Differences of approximately .1 were observed.

Table 11 presents average absolute errors for each parameter. The cell values are simple averages of the five calibrations contained in each. The marginal values are simple averages of the cell values. Relatively consistent decreases in the amounts of a-parameter error were apparent with increasing test length and calibration group size. These decreases were probably due to decreases in bias observed earlier because only minor differences were observed in correlation.

Table 11. Absolute Parameter Error
Basic Data Set--Randomly Sampled Examinees

Parameter	Sample Size	Test Length 20	35	50	65	Average
a	500	.839	.642	.491	.455	.607
	1000	.775	.531	.450	.472	.557
	2000	.841	.499	.404	.419	.541
	Average	.818	.557	.448	.449	
b	500	.314	.298	.285	.262	.290
	1000	.239	.271	.275	.247	.258
	2000	.316	.196	.209	.233	.238
	Average	.290	.255	.256	.247	
c	500	.136	.128	.108	.110	.120
	1000	.128	.111	.095	.085	.105
	2000	.146	.098	.092	.096	.108
	Average	.137	.112	.098	.097	

Intuitively, these errors appear quite large because an a value of .8 is considered adequate for adaptive testing, and an average error this large was observed in the first column.

The second section of Table 11 shows slight and inconsistent decreases in absolute error of the b parameters with increasing test length and calibration group size. The decreases were somewhat more consistent with increasing calibration group size; with the exception of the 20-item test length, absolute errors decreased with increased sample size.

Errors in the c parameters generally decreased with increasing test length and group size. This trend appeared to be somewhat more consistent relative to group size than to test length. Noting that an average c parameter is approximately .2, the errors observed in Table 10 typically exceeded half this amount and seemed quite large.

Table 12 presents root-mean-square errors of estimate for the item parameters. Root-mean-square error can be interpreted in a manner similar to absolute error. The marginal averages in Table 11 were computed as the square root of the mean of the squares in

Table 12. Root-Mean-Square Parameter Error
Basic Data Set--Randomly Sampled Examinees

Parameter	Sample Size	Test Length				Average
		20	35	50	65	
a	500	.710	.522	.368	.359	.510
	1000	.680	.430	.341	.344	.469
	2000	.735	.422	.305	.295	.474
	Average	.709	.460	.339	.333	
b	500	.242	.239	.212	.196	.223
	1000	.195	.203	.202	.185	.196
	2000	.261	.155	.156	.163	.189
	Average	.234	.202	.191	.182	
c	500	.108	.101	.083	.080	.094
	1000	.103	.088	.074	.066	.084
	2000	.122	.074	.067	.071	.087
	Average	.112	.089	.075	.072	

the corresponding rows and columns. Essentially the same observations made regarding the absolute error can be made here regarding the root-mean-square errors.

Characteristics of Asymptotic Ability Estimates

Table 13 presents the average absolute error of estimate of ability that would be obtained if the calibrated items were administered an infinite number of times to an infinitely large standard normal population of examinees and were scored using the estimated parameters. Entries corresponding to the 12 cells are simple averages of this error obtained with five different sets of items. These errors are unlike the absolute errors discussed in the previous section in that they refer to asymptotic errors in the estimation of ability and not to errors in the item parameters themselves.

The absolute errors, presented in Table 13, consistently decreased as the test lengths increased and, except for one inconsistent cell, as calibration group size increased. The unit of these errors is the same as the standard theta metric and some comparison can be made with absolute errors in the b parameters presented in

Table 13. Absolute Asymptotic Ability Error
Basic Data Set--Randomly Sampled Examinees

Sample Size	Test Length 20	35	50	65	Average
500	.170	.140	.104	.107	.130
1000	.123	.102	.101	.093	.105
2000	.157	.093	.085	.086	.105
Average	.150	.112	.097	.095	

Table 11. The errors in the asymptotic ability estimates were somewhat smaller than those observed with the b parameters. This is probably due to an averaging effect across items. An important feature to note, however, is that these errors did not reach zero as test length reached infinity.

Root-mean-square errors of asymptotic ability estimates are presented in Table 14. Marginal values in this table were computed as the square root of the mean of the squared entries in the corresponding rows and columns. All of the same conclusions drawn from the previous table can be drawn from this one; with the exception of the lower left cell, all errors decreased with increasing test length and increasing calibration group size.

Table 14. Root-Mean-Square Asymptotic Ability Error
Basic Data Set--Randomly Sampled Examinees

Sample Size	Test Length 20	35	50	65	Average
500	.222	.172	.120	.129	.166
1000	.180	.156	.119	.112	.144
2000	.229	.118	.102	.102	.148
Average	.212	.151	.114	.115	

Efficiency of Ability Estimation

Table 15 shows relative efficiencies of calibration. Entries for each of the 12 cells were computed by first obtaining the average item information in each calibration sample of the cell, summing

Table 15. Relative Efficiency
Basic Data Set--Randomly Sampled Examinees

Sample Size	Test Length 20	35	50	65	Average
500	.843	.866	.899	.927	.884
1000	.863	.894	.916	.943	.904
2000	.818	.911	.943	.952	.906
Average	.841	.890	.919	.941	

the information obtained using the estimated parameters and using the true parameters, and then dividing the sum obtained from the estimated parameters by the sum obtained from the true parameters. The marginal efficiencies were computed as the simple average of the corresponding row or column efficiencies. Average item information was used as a starting point instead of test information to avoid implicitly weighting the constituents of the row averages by the length of the tests.

Efficiencies ranged from a low of .818 to a high of .952. These efficiency values can be interpreted in an absolute sense; they can be thought of in terms of effective numbers of items. If, for example, a 100-item test were composed of items calibrated in sets of 65 items administered to 2,000 examinees, the ability estimation capacity of the test would be about the same as if 95 items with true parameters were administered. If a test comprised of 100 items calibrated in sets of 20 on 500 examinees were used, this would be equivalent to an 84-item test using true parameters. This last test discussed would require .952/.843 = 1.12 times as many or 12% more items than the first test to achieve the same measurement precision.

With the exception of the lower left cell, all efficiencies consistently increased with increasing test length and calibration group size. More interesting than this qualitative evaluation, however, is the observation that an increase in test length produced a relatively larger change in efficiency than did calibration group size. Slightly more than tripling the test length from 20 to 65 items produced a change in efficiency of 11.9% (.941/.841 = 1.119). Quadrupling the calibration group size from 500 to 2000 examinees resulted in an increase of only 2.5%, less than one-fourth the increase observed from tripling the test length. The data from the randomly selected examinees thus suggest that test length is relatively more important than calibration group size in determining the efficiency of calibration, at least at test lengths and sample sizes evaluated here.

Systematically Sampled Examinees

Fidelity of Parameter Estimation

Table 16 presents the parameter bias statistics for item parameters calibrated on the systematically sampled examinees. The first section presents bias of the a parameters. As was observed with the randomly sampled examinees, the bias dropped as test length increased and exhibited no definite trend with calibration group size. All marginal bias values were about .10 units less than those observed with the randomly sampled examinees. This trend continued even as the bias values dropped below zero and became negative.

Table 16. Item Parameter Bias
Basic Data Set--Systematically Sampled Examinees

Parameter	Sample Size	Test Length 20	35	50	65	Average
a	500	.504	.074	.008	-.105	.120
	1000	.478	.184	.021	-.111	.143
	2000	.462	.223	.017	-.084	.155
	Average	.481	.160	.015	-.100	
b	500	.090	.298	.207	.151	.187
	1000	.186	.214	.045	.141	.147
	2000	.045	.073	-.067	.175	.057
	Average	.107	.195	.062	.156	
c	500	.042	-.001	.013	-.024	.007
	1000	.029	.007	-.013	-.021	.001
	2000	.026	.009	-.022	-.009	.001
	Average	.032	.005	-.007	-.018	

Bias in the b parameters exhibited no obvious trend with increasing test length. This is different from the random-sampling case which exhibited a slight decrease. The same slight decrease with respect to calibration group size was again observed, however. The range in bias of the b parameters was somewhat larger in these samples. Where the range was from .066 to .155 in the random samples, the range was from -.067 to .298 in these samples.

Bias values of the *c* parameters also had a wider range in these samples. Where the random samples had bias values ranging from -.004 to .033, these samples had values ranging from -.022 to .042. The slight trend toward less bias observed in the random samples had an analog in the systematic samples; the trend could better be described as a trend toward more negative bias, however. Again, no consistent trend was observed with respect to calibration group size.

Table 17 presents the average correlations between true and estimated parameters for the systematically sampled calibration groups. As with the randomly sampled groups, a slight but inconsistent increasing trend of the *a*-parameter correlations with respect to test length was observed. No trend with respect to calibration group size was obvious, however. The overall magnitude of the *a*-parameter correlations in the systematically sampled groups was slightly lower than those observed in the randomly sampled groups.

Table 17. Parameter Correlations
Basic Data Set--Systematically Sampled Examinees

Parameter	Sample Size	Test Length 20	35	50	65	Average
a	500	.560	.582	.562	.463	.543
	1000	.204	.609	.582	.579	.508
	2000	.355	.601	.709	.664	.596
	Average	.383	.597	.622	.574	
b	500	.972	.976	.987	.979	.979
	1000	.984	.987	.986	.985	.986
	2000	.982	.985	.990	.989	.987
	Average	.980	.983	.988	.985	
c	500	.437	.360	.396	.381	.394
	1000	.448	.438	.416	.396	.425
	2000	.372	.375	.421	.519	.424
	Average	.420	.391	.411	.434	

The *b*-parameter correlations exhibited slight increasing trends with respect to test length and calibration group size. As was observed in the randomly sampled groups, these trends were inconsistent.

The magnitudes of the correlations were slightly lower in the systematically sampled groups.

No trends were apparent in the *c*-parameter correlations. Unlike those of the random samples, no notable increase was observed at a test length of 35 or a sample size of 1000. The magnitudes of the *c*-parameter correlations were somewhat lower here than those observed in the random samples.

Average absolute errors of the item parameters for the systematically sampled groups are presented in Table 18. A decreasing trend in *a*-parameter errors with respect to test length was apparent but was not particularly consistent. No trend was obvious in the *a*-parameter errors with respect to calibration group size. The magnitudes of the errors observed here were about the same as those observed in the randomly sampled groups.

Table 18. Absolute Parameter Error
Basic Data Set--Systematically Sampled Examinees

Parameter	Sample Size	Test Length 20	35	50	65	Average
a	500	.772	.535	.500	.549	.589
	1000	.911	.545	.471	.468	.599
	2000	.758	.574	.402	.433	.542
	Average	.814	.551	.457	.484	
b	500	.519	.577	.451	.496	.511
	1000	.500	.466	.322	.435	.431
	2000	.380	.437	.325	.440	.396
	Average	.466	.493	.366	.457	
c	500	.166	.134	.120	.123	.136
	1000	.149	.121	.116	.123	.127
	2000	.153	.134	.124	.107	.130
	Average	.156	.130	.120	.118	

The *b* parameters exhibited no trend in absolute error with respect to test length. A consistent decrease in error with respect to increasing sample size was observed. This supports the findings with the randomly sampled groups where no trend was observed with respect to test length but a slight trend was observed with respect to group size. The magnitudes of the errors were greater here than in the randomly sampled groups.

The c-parameter errors showed a relatively consistent decreasing trend with respect to test length but no consistent trend with respect to sample size. These findings are similar to those of the randomly sampled groups except that a slight trend with respect to group size was observed there. Magnitudes of the errors were slightly higher in the systematically sampled groups.

Table 19 presents the root-mean-square errors of estimate for the three parameters. As was the case in analysis of the randomly sampled groups, essentially the same observations made regarding the absolute error can be made regarding the root-mean-square error.

Table 19. Root-Mean-Square Parameter Error
Basic Data Set--Systematically Sampled Examinees

Parameter	Sample Size	Test Length 20	35	50	65	Average
a	500	.668	.419	.373	.388	.477
	1000	.772	.417	.338	.341	.500
	2000	.687	.464	.293	.307	.465
	Average	.710	.434	.336	.347	
b	500	.425	.530	.395	.405	.442
	1000	.411	.439	.251	.350	.370
	2000	.335	.377	.261	.397	.346
	Average	.392	.453	.309	.385	
c	500	.141	.107	.096	.098	.112
	1000	.129	.099	.094	.103	.107
	2000	.131	.112	.101	.090	.109
	Average	.134	.106	.097	.097	

Characteristics of Asymptotic Ability Estimates

Table 20 presents the absolute errors of asymptotic ability estimates for items calibrated using systematically sampled groups. Unlike the corresponding table for the randomly sampled groups, no consistent trends with respect to test length or sample size were observed. The magnitudes of the errors were consistently larger, however. Absolute errors in the randomly sampled groups ranged from .085 to .170; in the systematically sampled groups they ranged from .124 to .346.

Table 20. Absolute Asymptotic Ability Error
Basic Data Set--Systematically Sampled Examinees

Sample Size	Test Length 20	35	50	65	Average
500	.320	.336	.227	.266	.287
1000	.346	.313	.124	.215	.249
2000	.225	.263	.137	.293	.229
Average	.297	.304	.163	.258	

Similar observations can be made for the root-mean-square errors presented in Table 21. No definite trends were apparent and the magnitude of the errors was larger than in the randomly sampled groups. Root-mean-square errors ranged from .102 to .229 in the randomly sampled groups; in the systematically sampled groups they ranged from .158 to .466.

Table 21. Root-Mean-Square Asymptotic Ability Error
Basic Data Set--Systematically Sampled Examinees

Sample Size	Test Length 20	35	50	65	Average
500	.366	.434	.303	.330	.362
1000	.456	.349	.158	.249	.327
2000	.288	.305	.179	.346	.286
Average	.381	.367	.223	.311	

Efficiency of Ability Estimation

Table 22 presents the efficiencies of the items calibrated in the systematically sampled groups. The general trends observed in the randomly sampled groups were again observed here. In these groups, tripling the test length increased the calibration efficiency by 9.8%, and quadrupling the calibration sample size only increased the efficiency by 3.2%. Although the differences were not as pronounced, these results corroborated the earlier ones, suggesting that test length is more important than group size in improving calibration efficiency.

Table 22. Relative Efficiency
Basic Data Set--Systematically Sampled Examinees

Sample Size	Test Length				Average
	20	35	50	65	
500	.851	.851	.904	.901	.877
1000	.797	.877	.910	.930	.879
2000	.870	.884	.930	.934	.905
Average	.839	.871	.915	.922	

The magnitudes of the efficiencies were approximately equal in the two conditions. Efficiencies of the randomly sampled groups ranged from .818 to .952. Efficiencies of the systematically sampled groups ranged from .797 to .934. It is difficult to say whether the slight superiority of the randomly sampled groups was due to more appropriate ability distributions, all being standard normal, or simply to sampling error.

Selected Examinees

Fidelity of Parameter Estimation

Table 23 presents bias statistics for the parameters of items calibrated on selected samples of examinees. All samples contained 1,000 examinees, so only four cells and their row average are presented in the table. Bias in the *a* parameters ranged from -.283 to -.416. A consistent decreasing trend with increasing test length was obvious. The bias progressed to a value more negative than observed in either of the calibration groups discussed above.

Table 23. Item Parameter Bias
Basic Data Set--Selected Examinees

Parameter	Test Length 20	35	50	65	Average
a	.416	-.031	-.164	-.283	-.015
b	-.213	-.459	-.377	-.464	-.378
c	.145	.128	.095	.075	.111

The *b* parameters had a consistent negative bias. This was undoubtedly due to the fact that the selected population had higher ability than the standard (i.e., 0,1) population assumed by the calibration procedure. No trend with respect to test length was observed.

Bias in the *c* parameters consistently decreased with increasing test length. The bias was considerably higher than that observed in corresponding tables for the other samples. Average bias for the random and systematic samples of 1,000 examinees were .012 and .001. Both were much lower than the .111 observed here.

Table 24 presents correlations between the true and estimated parameters for the selected-examinee samples. No consistent trend was observed in the *a*-parameter correlations with respect to test length but the correlations generally rose with increasing test length. The correlations were somewhat lower than those observed in the corresponding rows of previous tables, however.

Table 24. Parameter Correlations
Basic Data Set--Selected Examinees

Parameter	Test Length 20	35	50	65	Average
a	.349	.408	.595	.502	.469
b	.983	.973	.979	.975	.978
c	.335	.323	.319	.388	.342

The *b*-parameter correlations exhibited no trend with respect to test length. Their average value of .978 was slightly lower than those of .989 and .986 observed for the randomly and systematically selected groups, respectively.

No trend was apparent in the *c*-parameter correlations, either. Their average of .342 was lower than the values of .541 and .425 observed in the two previous calibration groups. This should be expected, however, because the selected group (in which only the most able two-thirds of the examinees were selected) provided few of the low-ability examinees needed to accurately estimate the *c* parameters.

Table 25 presents average absolute errors of the item parameters. The *a*-parameter errors generally decreased as test length increased. The magnitude of the row average was slightly higher than the corresponding row averages for the randomly or systematically sampled groups.

Table 25. Absolute Parameter Error
Basic Data Set--Selected Examinees

	Test Length				
Parameter	20	35	50	65	Average
a	.800	.655	.536	.560	.638
b	.553	.705	.624	.716	.649
c	.202	.186	.156	.138	.170

Errors in the *b* parameters showed no consistent trend with respect to test length. The row average, .649, was considerably higher than the row averages for the randomly or systematically sampled groups, respectively .258 and .431.

The *c*-parameter errors showed a decreasing trend with respect to test length ranging from .202 to .138. The row average, .170 was somewhat higher than those of corresponding rows in previous tables.

Table 26 presents root-mean-square errors for the item parameters in the selected examinee groups. Again, the results closely parallel those of the absolute errors.

Table 26. Root-Mean-Square Parameter Error
Basic Data Set--Selected Examinees

	Test Length				
Parameter	20	35	50	65	Average
a	.658	.510	.403	.411	.506
b	.459	.578	.500	.568	.529
c	.181	.158	.125	.111	.146

Characteristics of Asymptotic Ability Estimates

Table 27 presents absolute and root-mean-square asymptotic ability-estimation errors. Absolute errors showed no trend with respect to test length. The average of the row, .580, was considerably larger than the averages of .105 and .249 observed in corresponding earlier tables.

Table 27. Asymptotic Ability Error
Basic Data Set--Selected Examinees

	Test Length				
Error	20	35	50	65	Average
Absolute	.499	.633	.558	.630	.580
Root-Mean-Square	.591	.744	.642	.754	.686

The root-mean-square errors showed an identical lack of trend with respect to test length. Similarly, the row average of .686 was considerably larger than the row averages of .144 and .327 observed earlier.

Efficiency of Ability Estimation

Calibration efficiencies obtained in the selected samples of examinees are presented in Table 28. The usual trend with respect to test length, observed with other statistics, was again observed. The average efficiency, .823, was somewhat lower than the corresponding

Table 28. Relative Efficiency
Basic Data Set--Selected Examinees

Test Length 20	35	50	65	Average
.719	.818	.865	.889	.823

efficiencies of .904 and .879 observed earlier. This lowered efficiency cannot be attributed to any particular item parameter because all three were less precisely estimated in this calibration sample than in the two discussed previously. It was probably due to the combined effects of poorly estimated c parameters, caused by a paucity of low-ability examinees, and fewer appropriate items for ability estimation at the higher ability levels encountered. This latter effect is due to limitations of the item pool used but these limitations were imposed to reflect reality, and thus the same effect in live-examinee item calibrations would be expected.

Conclusions

Three general conclusions and an observation can be made from the data presented in this section. First, the parameter correlation data were, in general, supportive of other studies investigating the calibration effectiveness of OGIVIA. The b parameters were very well estimated and the a and c parameters were less well estimated. The a parameters were estimated somewhat better than the c parameters, but the difference was not overwhelming.

The second conclusion is that test length is relatively more important to calibration effectiveness than is sample size, at least at the test lengths and sample sizes investigated here. This conclusion is mildly supported by the fidelity of estimation data but its strongest support comes from the efficiency analyses. The efficiency analyses suggested that increases in test length are at least three to four times as effective in improving calibration efficiency as proportionate increases in calibration sample sizes. Given that total testing time required to calibrate a set of items is proportional to the number of items multiplied by the number of examinees, this finding suggests that, if sufficient items exist, larger numbers of items should be calibrated on smaller samples if available total testing time is short.

The third conclusion is that there appears to be little difference in calibration efficiency as a function of random versus systematic sampling of examinees but a large difference between these and selected samples of examinees (as defined here). Although some differences were observed between random and systematic samples in the fidelity analysis, differences in the efficiencies were trivial and probably due to sampling error. Efficiencies observed in the selected samples were noticeably lower, however, and were probably due to a lack of low-ability examinees for c parameter estimation and to a distribution of abilities slightly less estimable with available items.

In addition to these conclusions, the parameter bias statistics presented in Tables 9, 16, and 23 suggest that OGIVIA tends to overestimate a parameters at short test lengths. Since the test lengths used to evaluate the real ASVAB data ranged from 20 to 35 items, and since OGIVIA was one of the estimation methods used, the average a value of 1.6 used to generate items for the simulations may have been too high. As can be seen from Tables 9, 16, and 23, the amount by which the a parameters are overestimated depends on the method by which subjects are selected and ranges from an overestimate of .4 units to .6 units for 20-item tests and from an overestimate of from .3 down to a nearly zero underestimate for 35 items. It is difficult to determine the extent to which the value used, 1.6, was biased but the fact that it was probably slightly high should be kept in mind.

IV. LINKING WHEN EXAMINEES ARE RANDOMLY SAMPLED

Linking sets of items administered to randomly sampled examinees presented the simplest linking environment investigated in this research. In this situation, the equivalent-groups, anchor-group, and anchor-test methods were all reasonable choices. Given the added assumption that items were randomly assigned to forms, usually an easy assumption to satisfy, the equivalent-tests method was also an acceptable method.

The basic data set containing randomly sampled examinees was used for this portion of the research. Although all four linking paradigms were conceptually reasonable to apply, only the equivalent-groups and equivalent-tests methods were evaluated. The anchor-group and anchor-test linking methods were not evaluated using this data set where examinees were randomly sampled from a single population. This deletion was done purely for efficiency of analysis. Since these methods do not assume randomly sampled examinees, it was reasonable to expect that data from the systematic examinee samples would yield sufficient data for comparison. Given the reasonableness of this expectation and the extensive amount of computer time required to analyze those methods, a decision was made not to perform this essentially duplicate analysis.

Equivalence Methods

The equivalent-groups and equivalent-tests methods are essentially the same in terms of the data required. The differences between them stem from the different assumptions invoked in obtaining the transformation parameters. The two methods have thus, for purposes of this report, been combined into one section. Although they are discussed as separate methods, they share common tables.

Procedure

Equivalent groups. Conceptually, equivalent-groups linking is accomplished by finding transformation constants which, when applied to the a and b parameters, will make the mean and variance of ability in each group equivalent. Two transformation constants are required to accomplish this. Given that the constants are to be applied in the form:

$$a = dk \qquad [14]$$

and

$$b = (e-m)/k \qquad [15]$$

where $\underline{a}$ and $\underline{b}$ are the parameters on the "equivalent" metric and $\underline{d}$ and $\underline{e}$ are the parameters on the unlinked metric, one set of constants that will result in a common metric with a mean of zero and variance of one is:

$$k = \sigma_{\Gamma} \qquad [16]$$

and

$$m = \mu_{\Gamma} \qquad [17]$$

where μ_{Γ} and σ_{Γ} are, respectively, the mean and standard deviation of ability estimates in the unlinked groups. These values may be readily verified by noting that a satisfactory transformation must satisfy the equation:

$$a(\theta - b) = d(\Gamma - e) \qquad [18]$$

If $\underline{a}$ and $\underline{b}$ given in Equations 14 and 15 are substituted into Equation 18, gamma can be expressed as a function of $\underline{k}$, $\underline{m}$, and theta:

$$= k\theta + m \qquad [19]$$

Given that theta is to be distributed with mean zero and variance one, the constants $\underline{k}$ and $\underline{m}$ are obviously the standard deviation and the mean of gamma. Thus, the constants in the equivalent-groups method are simply the mean and standard deviation of the abilities in the unlinked groups.

In practice, true abilities are not available, however, and they must be estimated. If errors of measurement are equivalent in each group or adequately compensated for, equivalent-groups linking may be accomplished using ability estimates. There are, however, several such estimates that may be used. Four methods of estimating ability were investigated including two Bayesian and two maximum-likelihood methods. In addition to simple means and standard deviations of these estimates, robust estimation procedures were applied to the maximum-likelihood estimates. This resulted in six methods for determining the equivalent-groups transformation constants.

The program OGIVIA uses a modal Bayesian estimate with a standard-normal prior ability assumption. The estimates provided by OGIVIA were based on an early stage of the program which did not use the final item parameter estimates. Proceeding in the spirit of OGIVIA but using better parameter estimates, modal Bayesian ability estimates assuming a standard-normal prior were obtained by solving the following equation for theta:

$$\hat{\theta} = 1.7 \sum_g a_g \exp(x_g) \left[\frac{u_g}{c_g + \exp(x_g)} - (1.0 + \exp(x_g))^{-1} \right] \qquad [20]$$

where $u_g = 1$ if the item is answered correctly

$= 0$ otherwise

and $x_g = 1.7\ a_g(\hat{\theta} - b_g)$

The Bayesian estimation procedure assuming a normal prior implicitly regresses the estimates at finite test lengths. The practical effect of this on linking is to bias the linking constants. The second estimation procedure incorporated an attempt to correct for this regression by progressing the estimation by an amount equivalent to the suspected regression. This adjustment was accomplished by using the Bayesian posterior variance estimate obtained from Equation 21 and the Bayesian ability estimate obtained from Equation 20 as prescribed in Equation 22.

$$\theta_B^2 = \left\{ -1 + 2.89 \sum_g a_g^2 \exp(x_g) \left[\frac{c_g}{(c_g + \exp(x_g))^2} - (1.0 + \exp(x_g))^{-2} \right] \right\}^{-1} \qquad [21]$$

$$\hat{\theta}_{B_{Pro}} = \hat{\theta}_B\ (1 - \sigma_B^2)^{-1/2} \qquad [22]$$

Another procedure to ameliorate the Bayesian regression is to use a maximum-likelihood estimation procedure instead of a Bayesian one. The maximum-likelihood procedure attempts to be unbiased and does not regress the ability estimates. It has problems, however, in that it tends to make some extreme estimates when the test length is finite. Individuals answering all items correctly or less than a chance number correctly receive infinite ability estimates. Such estimates, in turn, cause some difficulty in calculation of means and variances of the ability estimates. Maximum-likelihood estimation was used as the third estimation procedure. In most cases, these estimates were obtained by finding the root in theta of Equation 23:

$$\sum_g a_g \exp(x_g) \left[\frac{u_g}{c_g + \exp(x_g)} - (1.0 + \exp(x_g))^{-1} \right] = 0 \qquad [23]$$

In cases where the estimates were beyond plus or minus 3.5, the estimates were artificially bounded at those values.

The Bayesian procedure was corrected for regression. An attempt was made to correct the maximum-likelihood procedure for erring toward the extreme. This was accomplished by applying the squared standard error of estimate obtained from Equation 24 to the ability estimate obtained from Equation 23 by the method prescribed in Equation 25.

$$\sigma^2 = \left\{ 2.89 \sum_g a_g^2 \exp(x_g) \left[\frac{c_g}{(c_g + \exp(x_g))^2} - (1.0 + \exp(x_g))^{-2} \right] \right\}^{-1} \qquad [24]$$

$$\hat{\theta}_{Reg} = ((\hat{\theta} - \bar{\hat{\theta}})\ (1 - \sigma^2)^{1/2}) + \hat{\theta} \qquad [25]$$

Truncation of the ability estimates at plus and minus 3.5 was one method of dealing with extreme ability estimates produced by the maximum-likelihood procedure. This method was somewhat arbitrary and still used a least-squares weighting scheme within the range. General procedures of robust estimation were available to deal with problems such as these. One of the most popular procedures was the AMT sine-transformation procedure (Andrews, Bickel, Hampel, Huber, Rogers, & Tukey, 1972; Wainer & Wright, 1980). In this procedure, the equation

$$\sum f[(\hat{\theta} - T)/S] = 0 \qquad [26]$$

is solved for T and S where T is the robust estimate of location, S is the median absolute deviation from T divided by the constant 1.349, and

$$f[x] = \sin(x/2.1) \qquad \text{if } -6.597 < x \leq 6.597 \qquad [27]$$

and $f[x] = 0$ otherwise.

The procedure was iterated adjusting both T and S on each iteration until T stabilized within 0.001.

This robust estimation procedure was applied to the maximum-likelihood estimates and the regressed-maximum-likelihood estimates obtained above to produce the fifth and sixth methods of estimating the mean and standard deviation of ability. Unlike the first four methods, the robust techniques were not methods of estimating ability but rather methods of obtaining means and standard deviations of estimates. The means and standard deviations were the only elements used for linking, however, and these robust procedures thus produced two more methods of equivalent-groups linking. It should be noted that the robust techniques were applied to the truncated maximum-likelihood estimates and not to estimates permitting infinite values.

Equivalent tests. The equivalent-tests method assumes that the item parameter distributions of the tests being linked are equivalent. Linking, under this assumption, is accomplished by setting the a and b parameters to common values in each of the tests. Practically, these values can be any values desired. To aid in interpretation of the fidelity and asymptotic characteristic statistics, these common values were set to the true means obtained in the simulation reported in the design section of this report, 1.586 and 0.227 for a and b, respectively. This was accomplished by computing transformation parameters k and m as follows:

$$k = 1.586/\mu_d \qquad [28]$$

$$m = (\mu_e - 0.360)/\mu_d \qquad [29]$$

where μ_d and μ_e are the means of the a and b parameter estimates in each test prior to linking.

Results

The magnitude of the amount of data generated by this project made it unreasonable to present all analyses in the body of this report. To meaningfully present the analyses done, individual tables are presented in the Technical Appendix and summary tables are presented here in the text. For the homogeneous linking evaluation in which linking was done separately in each of the 12 cells, 12 individual tables are presented for each of the three classes of analyses in the Technical Appendix. One composite table is presented in the body of the report for each class of analysis. For the heterogeneous linking evaluation where five replications pooling 20 items from each cell were done, five individual tables for each class of analysis are presented in the Technical Appendix, and one is presented in the body of the report.

Fidelity of parameter estimation. Table 29 presents fidelity-of-parameter-estimation statistics for eight linking methods in the homogeneous condition. The first six methods correspond to different methods of determining the linking constants within the equivalent-groups method. The seventh is the equivalent-tests linking method. The "no-linking" method is included as a baseline of comparison in which

Table 29. Item Parameter Error--Equivalence Methods
Homogeneous Condition Using Randomly Sampled Examinees

	True		Bias in		Absolute	RMS	
Method	Mean	SD	Mean	SD	Error	Error	R
Bayesian							
a	1.591	.482	-.020	.018	.344	.469	.581
b	.221	1.329	.088	.311	.293	.425	.987
Progressed Bayes							
a	1.591	.482	.041	.036	.359	.484	.581
b	.221	1.329	.072	.250	.255	.370	.987
Max. Likelihood							
a	1.591	.482	.344	.125	.527	.693	.576
b	.221	1.329	.023	.019	.171	.234	.987
Regressed M.L.							
a	1.591	.482	.223	.088	.454	.605	.576
b	.221	1.329	.036	.106	.190	.264	.987
Robust M.L.							
a	1.591	.482	.263	.112	.473	.616	.578
b	.221	1.329	.043	.076	.186	.271	.986
Rob. Reg. M.L.							
a	1.591	.482	.202	.093	.435	.572	.579
b	.221	1.329	.048	.121	.198	.295	.986
Equivalent Tests							
a	1.591	.482	-.006	.015	.337	.456	.577
b	.221	1.329	.006	.275	.358	.487	.974
No Linking							
a	1.591	.482	.236	.091	.453	.596	.581
b	.221	1.329	.110	.087	.198	.268	.987

the parameters were taken directly from OGIVIA with no explicit transformation. In fact, this procedure approximates an equivalent-groups linking method because OGIVIA, in an early stage of calibration, sets its best estimates of the mean and variance of ability to zero and one.

The first column presents the means of the true $\underline{a}$ and $\underline{b}$ parameters for all cells in the data set. To compute the values in the first column, means of parameters for all items in a cell were computed for that cell. This included all items in the five calibration groups. The mean of these 12 cell means was then computed for the entry in Table 29. The means of the $\underline{a}$ and $\underline{b}$ parameters, 1.591 and .221, were quite close to the means obtained in independent simulation (discussed with the analysis of the basic data sets) of 1.586 and .227.

The standard deviations presented in column two were computed as the square root of the mean variance averaged in the same manner as the means of column one. The averages of 0.482 and 1.329 were, again, very close to those obtained in simulation, 0.488 and 1.338.

Biases presented in columns three and four were computed as the linked value minus the true value for both means and standard deviations. Mean biases were computed for items in each of the 12 cells. Table 29 presents the means of these 12 cell means.

Absolute error was computed for each cell as the mean of the absolute deviations of linked from true item parameters for all items in a cell. Table 29 presents the simple average of these means over all 12 cells.

Root-mean-square error was calculated for each cell in a manner similar to that of absolute error. The squared deviations were averaged (rather than the absolute deviations), and the square root of the resultant mean was taken. The RMS error presented in Table 29 is the square root of the mean of the squared individual cell values.

Correlations between true and estimated parameters were computed in each of the 12 cells. An r-to-z average of the cell values was then taken for each entry in Table 29.

Compared in terms of bias, the equivalent-tests method of linking produced estimates closest in mean $\underline{a}$ and mean $\underline{b}$. It also produced estimates with the least bias in standard deviation of $\underline{a}$. Several methods had superior estimates in terms of standard deviations of the $\underline{b}$ parameters, however.

The equivalent-tests method was again superior when absolute error in the $\underline{a}$ parameters of the various methods was considered. Equivalent-groups methods based on either of the Bayesian procedures were nearly as good. When *b* parameters were considered, the maximum-likelihood procedures appeared to produce less absolute error than the other methods.

Root-mean-square error comparisons produced the same findings: the equivalent-tests method was superior in estimation of the *a* parameters with the Bayesian equivalent-groups methods close behind. The maximum-likelihood equivalent-groups methods produced the best estimates of the *b* parameters.

Correlational analyses showed the Bayesian and no-linking procedures to produce the best-linked *a* parameters. The maximum-likelihood procedures did nearly as well. The equivalent-tests method produced *a*-parameter correlations about as high as those of the maximum-likelihood methods. The *b*-parameter correlations were nearly constant at .986 to .987 for all but the equivalent-tests method, which produced a correlation of only .974.

Table 30 presents fidelity statistics for the heterogeneous linking condition containing pooled results of five replications sampling 20 items from each cell. Again, all entries are summary statistics of several individual tables contained in the Technical Appendix. In this case each entry represents pooled results of five replications rather than of 12 cells. The columns of the table all correspond to those of Table 29, and the pooling, in each case, was done in the same manner.

The means and standard deviations presented in the first two columns were again close to the true values found in the independent simulation. That they were slightly different is due to the fact that only the first 20 items in each calibration group were used for the heterogeneous analysis. Thus, less than half of the items included in the homogeneous analysis were used in this analysis.

The bias data in columns three and four presented essentially the same picture as the bias data in Table 29. Similarly, identical observations could be made regarding the absolute and root-mean-square error data of columns five and six. This similarity is more an artifact than a discovery, however, as neither the biases nor the errors are affected by composition of the item sets. The fact that they differ at all is due to fluctuations caused by item sampling.

The change in composition was expected to affect the correlations. Different test lengths and calibration group sizes do produce different biases in linking constants. The different biases shift items of the different cells differentially and this affects the correlations among the parameters. Marked changes from Table 29 occurred in Table 30. Where Table 29 showed *a*-parameter correlations closely clustered in value, the *a*-parameter correlations presented in Table 30 had a relatively wide range of values. Furthermore, the equivalent-tests method, which produced the lowest correlation in Table 29, produced the highest

Table 30. Item Parameter Error—Equivalence Methods
Heterogeneous Condition Using Randomly Sampled Examinees

Method	True Mean	True SD	Bias in Mean	Bias in SD	Absolute Error	RMS Error	R
Bayesian							
a	1.588	.490	-.014	.038	.348	.470	.580
b	.248	1.350	.090	.315	.295	.431	.983
Progressed Bayes							
a	1.588	.490	.047	.060	.363	.487	.577
b	.248	1.350	.073	.253	.258	.375	.984
Max. Likelihood							
a	1.588	.490	.350	.202	.532	.698	.529
b	.248	1.350	.020	.018	.175	.240	.985
Regressed M.L.							
a	1.588	.490	.229	.152	.459	.610	.535
b	.248	1.350	.035	.107	.194	.271	.985
Robust M.L.							
a	1.588	.490	.270	.157	.478	.622	.548
b	.248	1.350	.042	.078	.191	.279	.983
Rob. Reg. M.L.							
a	1.588	.490	.209	.130	.441	.577	.557
b	.248	1.350	.047	.125	.204	.303	.983
Equivalent Tests							
a	1.588	.490	.001	.032	.340	.459	.596
b	.248	1.350	.008	.277	.361	.491	.964
No Linking							
a	1.588	.490	.242	.144	.458	.600	.553
b	.248	1.350	.108	.086	.200	.273	.986

in Table 30. With the exception of this method, the a-parameter correlations were lower in Table 30 than in Table 29. The b-parameter correlations lost some of the uniformity they exhibited in Table 29 but the same general conclusions could be drawn. The equivalent-tests method was still inferior in terms of b-parameter correlations.

Characteristics of asymptotic ability estimates. Table 31 presents statistics descriptive of linking and calibration errors on asymptotic estimates of ability in the homogeneous condition. The

Table 31. Asymptotic Ability Estimates--Equivalence Methods
Homogeneous Condition Using Randomly Sampled Examinees

Method	Mean	SD	Absolute Error	RMS Error	R
Bayesian	.004	1.073	.064	.098	.999
Progressed Bayes	.001	1.035	.043	.072	.999
Max. Likelihood	-.002	.890	.100	.140	.998
Regressed M.L.	-.005	.945	.066	.100	.999
Robust M.L.	.002	.915	.079	.111	.999
Rob. Reg. M.L.	-.003	.944	.061	.088	.999
Equivalent Tests	-.086	1.066	.151	.209	.998
No Linking	.074	.934	.100	.125	.999

values in the table were compiled from corresponding values in 12 cells. The means and absolute errors in Table 31 represent simple averages of the cell values. The standard deviations and root-mean-square errors were computed as the square root of the mean squared values from the individual tables. The correlations were computed as the r-to-z average of the individual correlations.

The means, presented in the first column, were all fairly close to the true value of zero. The means produced by the six equivalent groups methods were all somewhat closer than the means produced by the equivalent-tests method or by no linking. The standard deviations were near the true value of 1.0 but were, typically, not as close as the means had been. The most deviant was the maximum-likelihood equivalent-groups procedure. The least deviant was the progressed-Bayesian equivalent-groups procedure.

Columns three and four present absolute and root-mean-square errors of the asymptotic estimates. The eight linking procedures ranked essentially the same in the two columns; the absolute errors produced a tie and the root-mean-square errors did not. The progressed-Bayesian equivalent-groups procedure produced the least error. The equivalent-tests procedure produced the most, more than the no-linking condition. Except for the equivalent-tests method, all methods (including no-linking) produced lower errors in asymptotic estimates than were produced by the unlinked individual calibrations summarized in Tables 13 and 14. Average values in those tables for absolute and root-mean-square error, respectively, were .113 and .153. The observation that error in the no-linking condition decreased was apparently due to a better averaging of parameter errors when all five calibration groups within a cell were combined.

The correlations between true and asymptotic ability estimates were so high as to be uninformative about linking adequacy of the various methods. All were within .002 of unity and, although the maximum-likelihood equivalent-groups and the equivalent-tests methods were slightly inferior, this difference may have been due to accentuation of trivial differences incurred in rounding.

Table 32 presents asymptotic error statistics for the heterogeneous condition. Again, all values are summary values and were prepared, in the same manner as Table 31, from five replications, each of which sampled 20 items from each of the 12 cells. The first two columns, those of the mean and standard deviation, were essentially unchanged from Table 31. The only difference was a slight tendency toward more extreme deviations of the standard deviations from 1.0. The two Bayesian methods were exceptions to this, in that they were slightly less deviant than in the homogeneous condition.

Table 32. Asymptotic Ability Estimates--Equivalence Methods
Heterogeneous Condition Using Randomly Sampled Examinees

Method	Mean	SD	Absolute Error	RMS Error	R
Bayesian	.006	1.064	.059	.084	.999
Progressed Bayes	.003	1.025	.037	.059	.999
Max. Likelihood	.002	.870	.108	.139	.999
Regressed M.L.	-.001	.927	.064	.089	.999
Robust M.L.	.004	.904	.081	.110	.999
Rob. Reg. M.L.	-.000	.933	.059	.085	.999
Equivalent Tests	-.087	1.075	.100	.143	.998
No Linking	.076	.919	.100	.123	.999

The absolute and root-mean-square errors showed some changes from the preceding table. The ordering of methods by the two statistics was not identical in Table 32. The Bayesian methods were still superior to all other methods. The equivalent-groups method improved to a point where it was nearly as good as no linking and, depending on the type of error, slightly better or slightly worse than the maximum-likelihood method.

The correlations presented in the fifth column were, again, particularly uninformative. Only one, that corresponding to the equivalent-tests method, showed any departure from the nearly perfect .999.

Efficiency of ability estimation. Table 33 presents efficiency data for the homogeneous linking condition. The first column contains the average item information produced in several ways. The first entry indicates the information available in the average item using true parameters. The second entry indicates information available using estimated parameters and (hypothetical) perfect linking. The remaining entries in the first column indicate information available from items using parameters linked in various ways.

Table 33. Efficiency Analysis--Equivalence Methods
Homogeneous Condition Using Randomly Sampled Examinees

Method	Average Item Information	Efficiency Relative to True Parameters	Efficiency Relative to Estimated Parameters
True Parameters	.319		
Est. Parameters	.287	.898	
Bayesian	.284	.888	.988
Progressed Bayes	.284	.888	.988
Max. Likelihood	.284	.889	.989
Regressed M.L.	.284	.888	.989
Robust M.L.	.284	.888	.989
Rob. Reg. M.L.	.284	.888	.988
Equivalent Tests	.276	.864	.962
No Linking	.284	.887	.988

Information from the true parameters was calculated separately in each of the individual calibration groups in each of the 12 cells using true parameters. The individual information values were then averaged to produce the value, .319, in Table 33. The information from the estimated parameters (the second entry) was obtained in the same way except that estimated parameters rather than true parameters were used. Since the computations were done within individual calibration groups, linking had no effect on the values.

The remaining values in the first column were obtained by pooling all items in each cell after the linking transformations were applied. The essential difference between these values and the information from the estimated parameters (i.e., the second entry) was that these values were obtained from a pool of all items in each cell rather than from each calibration group individually. The entries presented in Table 33 are simple averages of the corresponding entries in the 12 individual cell tables.

Efficiency relative to true parameters shown in column two was calculated directly from the values in column one of the table. Each value presented in column two is the corresponding value in column one divided by .319. Efficiency relative to estimated parameters was calculated similarly except that column one values were divided by .287. All columns in Table 33 present essentially the same data from a different viewpoint.

The efficiencies relative to estimated parameters provide data most directly relevant to comparisons of linking methods. These values can be interpreted as an index of linking efficiency. The information available from the estimated parameters calculated within individual calibration groups represents efficiency of calibration free of linking errors. Any degradation from that point, as items from several calibration groups are pooled, represents errors due to linking.

The efficiencies relative to estimated parameters suggest that there is very little difference among most linking methods in this condition. The notable exception is the equivalent-tests method. Where all other linking methods, including no-linking, had efficiencies of .988 or .989, the equivalent-tests method had a linking efficiency of only .962.

Table 34 presents efficiency statistics for the heterogeneous linking condition. All statistics were calculated in essentially the same manner as before. The primary difference was that the entries were computed as the average of five replication averages rather than as the average of 12 cell averages.

The information values for the true and estimated parameters changed very little from those of Table 33. The slight changes were

Table 34. Efficiency Analysis--Equivalence Methods
Heterogeneous Condition Using Randomly Sampled Examinees

Method	Average Item Information	Efficiency Relative to True Parameters	Efficiency Relative to Estimated Parameters
True Parameters	.317		
Est. Parameters	.285	.901	
Bayesian	.278	.875	.973
Progressed Bayes	.277	.876	.972
Max. Likelihood	.273	.861	.955
Regressed M.L.	.273	.863	.958
Robust M.L.	.276	.870	.965
Rob. Reg. M.L.	.276	.872	.967
Equivalent Tests	.269	.850	.944
No Linking	.274	.865	.960

due to the fact that only about half of the items on which Table 33 was based were used in computing the statistics of Table 34.

Marked changes in linking efficiency were noted, however. All methods, without exception, were less efficient in the heterogeneous condition. Differences among the methods were also more obvious. The two Bayesian methods were the most efficient. The robust maximum-likelihood procedures were next, followed by the no-linking method and the maximum-likelihood procedures. The equivalent-tests method was again the least efficient of all.

Table 35 presents linking efficiencies of the Bayesian equivalent-groups linking method for each of the 12 cells arranged by test length and sample size. The Bayesian procedure was singled out for this breakdown because it appeared, from data just presented, to be one of the best equivalent-groups linking procedures. Linking efficiency was chosen as the single statistic to be explored in this fashion because it seemed to best summarize the data to answer the question of which linking method allowed the best ability estimation. Individual cell entries in Table 35 were computed by taking the ratio

Table 35. Cellwise Efficiency Analysis
Bayesian Score--Randomly Sampled Examinees

Sample Size	Item Set Size 20	35	50	65	Average
500	.968	.991	.991	.959	.977
1000	.984	.990	.993	.996	.991
2000	.972	.993	.992	.996	.988
Average	.975	.991	.992	.984	

of the information values of the linked parameters to the information values of the estimated parameters calculated within individual calibrations. The marginal values presented are simple averages of the corresponding row and column values. They are not pooled values as were those in Tables 33 and 34 which were computed as ratios of averaged information values rather than averages of efficiencies.

No obvious relationships between linking efficiency and either test length or calibration sample size were observed. No trends were apparent, even in the marginal values. No interactions were apparent in the individual cell averages.

Table 36 presents a similar breakdown of the equivalent-tests method efficiencies. The marginal averages exhibited a definite increasing trend with increasing test length. This trend was not particularly consistent in the individual cell values, however. The

Table 36. Cellwise Efficiency Analysis
Equivalent Tests Randomly Sampled Examinees

Sample Size	Test Length 20	35	50	65	Average
500	.916	.985	.974	.966	.960
1000	.972	.930	.973	.986	.965
2000	.928	.961	.961	.982	.958
Average	.939	.959	.969	.978	

trend was apparent at sample sizes of 2,000 but not at 500 or 1,000. No relationship between efficiency and sample size was apparent in Table 36.

Discussion

Three sets of analyses have been presented. The fidelity analyses provided no conclusive evidence regarding which linking procedure was most effective. Data relevant to this were weak and conflicting. Methods most effective in linking a parameters were not the ones most effective in linking b parameters. There was no way to determine in any practical way whether a or b errors were more deleterious in regard to ability estimation.

The asymptotic estimation analysis was somewhat more helpful in that the joint effect of parameter errors on ability estimation could be observed. These data suggested that the two Bayesian linking procedures and the robust-regressed maximum-likelihood procedures were somewhat more effective than the others and that the equivalent-tests method was typically no better than the no-linking method.

Efficiency analyses suggested that whatever differences there were among the methods, they were quite small. Efficiency loss due to linking error was always less than loss due to calibration error, considerably less in some cases. In the worst case of linking error, information lost to linking was half as great as that lost to calibration. For the best linking methods information loss due to linking was 10% to 20% as large as that due to calibration, depending on the conditions.

Conclusions

Two general linking methods, the equivalent-groups and the equivalent tests methods, were evaluated and compared to each other and to a no-linking control method. These comparisons were done in both a homogeneous linking condition, where the items linked were calibrated in tests of the same length using examinee samples of equal size, and in a heterogeneous condition of mixed test lengths and sample sizes. Several conclusions can be drawn from these data.

First, the equivalent-groups methods were generally superior to the equivalent-tests method. In some analyses, reported in the fidelity of estimation section, the equivalent-tests method appeared to be superior. In the more readily interpretable asymptotic-estimate and efficiency analyses, the equivalent-tests method was consistently one of the poorer linking procedures.

Second, of the six equivalent-groups procedures evaluated, the ones based on the Bayesian scores appeared to be slightly superior to the others. This superiority was apparent only in the heterogeneous linking condition, however. In this condition a slight superiority was observed in the asymptotic estimation and efficiency analyses. Little difference among equivalent-groups procedures was observed in the homogeneous condition although the Bayesian methods had slightly less error in the asymptotic estimates than did some of the other procedures.

Third, it should be noted that the no-linking method worked reasonably well in these analyses. Although the other procedures produced slightly more efficient linking, relatively little efficiency would be lost, under the sampling characteristics present here, if the parameters were used as produced by OGIVIA with no explicit linking done.

Finally, although definite relationships between calibration efficiency and test length and sample size were shown in a previous section, no such relationships were found with respect to linking efficiency. This is counter-intuitive because all equivalence methods are dependent on sampling error which is dependent on sample size. Lack of any relationships may have been due to the fact that the range of sample sizes was too small to produce them. To the extent that this range covers the range of interest, however, the conclusion of no differences can reasonably be applied.

V. LINKING WHEN EXAMINEES ARE SYSTEMATICALLY SAMPLED

Linking with examinees systematically sampled represented an extreme case of violation of the assumption of random sampling essential to the equivalent-groups linking method. Only the equivalent-tests and the anchor methods were theoretically appropriate for this environment. Research reported in the previous section had shown the equivalent-groups method to be superior to the equivalent-tests method when the random-sampling assumption was satisfied. Thus, although it was not theoretically appropriate for this environment, the equivalent-groups method was evaluated to determine if it was practically acceptable.

The basic data set containing systematically sampled examinees was used for this portion of the research. For each calibration, an AFEES group was selected at random from the 65 available, and examinees were selected from that group. These data were then used in a manner similar to the data of the randomly sampled examinees.

Equivalence Methods

Procedure

The data used in this portion of the research differed from those reported in the previous section. The linking procedures used to implement the equivalent-groups and equivalent-tests methods did not differ, however. All six methods used for determining linking constants for the equivalent-groups method were again evaluated. The same linking transformation equations were again applied to both the equivalent-groups and the equivalent-tests methods.

Results

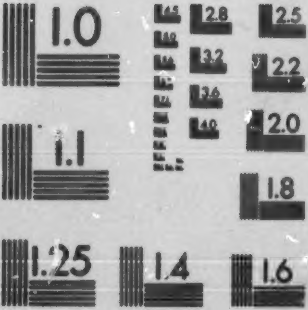
Fidelity of parameter estimation. Fidelity-of-estimation statistics for the homogeneous condition with systematically sampled examinees are presented in Table 37. True means and standard deviations, shown in the first two columns, were close to the population values. The mean of the $\underline{b}$ parameter, .262, was somewhat more deviant from the population value of .227 than the value observed in the previous section. All four values appeared to be well within the limits of sampling variation, however.

Bias in the estimated parameters is described in columns three and four. The Bayesian equivalent-groups methods tended to underestimate the $\underline{a}$ parameters. The maximum-likelihood procedures and the robust-maximum-likelihood procedures tended to overestimate the $\underline{a}$ parameters, although this was less the case with the non-robust

Table 37. Item Parameter Error--Equivalence Methods
Homogeneous Condition Using Systematically Sampled Examinees

	True		Bias in		Absolute	RMS	
Method	Mean	SD	Mean	SD	Error	Error	R
Bayesian							
a	1.588	.501	-.159	-.012	.374	.519	.533
b	.262	1.344	.173	.572	.568	.759	.971
Progressed Bayes							
a	1.588	.501	-.099	.008	.376	.517	.533
b	.262	1.344	.147	.495	.512	.682	.971
Max. Likelihood							
a	1.588	.501	.212	.111	.499	.674	.531
b	.262	1.344	.046	.188	.333	.423	.970
Regressed M.L.							
a	1.588	.501	.088	.073	.439	.596	.530
b	.262	1.344	.077	.295	.388	.493	.971
Robust M.L.							
a	1.588	.501	.194	.106	.470	.623	.529
b	.262	1.344	.054	.191	.334	.431	.970
Rob. Reg. M.L.							
a	1.588	.501	.107	.077	.425	.566	.531
b	.262	1.344	.079	.269	.375	.489	.971
Equivalent Tests							
a	1.588	.501	-.003	.034	.371	.510	.526
b	.262	1.344	-.035	.340	.417	.587	.971
No Linking							
a	1.588	.501	.139	.084	.450	.602	.533
b	.262	1.344	.130	.237	.364	.464	.971

regressed procedure. The equivalent-tests procedure produced little bias in the _a_ parameters. No-linking resulted in overestimation of _a_ parameters. Slight bias in the _b_-parameter means was produced by the two Bayesian procedures. The no-linking procedure produced a similiar amount of bias. The other procedures all produced somewhat less bias.


1.0
2.8
2.5
3.2
2.2
3.6
2.0
1.1
4.0
1.8
1.25
1.4
1.6

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS
STANDARD REFERENCE MATERIAL 1010a
(ANSI and ISO TEST CHART No. 2)

In terms of bias in parameter standard deviations, the Bayesian procedures produced the least bias for the a parameters. The maximum-likelihood procedures and the no-linking procedure produced the most bias in the a-parameter standard deviations. These observations essentially reversed when the b-parameter bias was considered; the Bayesian procedures produced the greatest bias, and the maximum-likelihood and no-linking procedures produced the least.

When the biases in columns three and four of Table 37 are compared to corresponding values for the randomly sampled examinees presented in Table 29, several things may be noted: The tendency of the maximum-likelihood and no-linking procedures to overestimate the a parameters was observed in both tables; biases in b-parameter means and a-parameter standard deviations were similiar in both tables; and the biases in the b-parameter standard deviations were somewhat larger in Table 37.

Absolute and root-mean-square errors of parameter estimation are presented in columns five and six of Table 37. The equivalent-tests method produced the least parameter error, evaluated by either statistic, for the a parameters. The two Bayesian methods were nearly as good, however. The maximum-likelihood and no-linking procedures produced the greatest amount of a-parameter error. The least b-parameter error was produced by the maximum-likelihood methods; the most was produced by the Bayesian methods.

Error in the a parameters observed in Table 37 was similar in magnitude to that observed in Table 29. Absolute errors of the a parameters ranged from .337 to .527 in Table 29; in Table 37 the comparable range was from .371 to .499. Error in the b parameters was somewhat greater in Table 37, however. Absolute errors of the b parameters ranged from .171 to .358 in Table 29; in Table 37 they ranged from .333 to .568.

Correlations between true and estimated a parameters, shown in column seven, were very similar for all linking methods. The Bayesian, the robust-regressed maximum-likelihood, and the no-linking procedures were best, with correlations of .533. The equivalent-tests method was worst, with a correlation of .526. Correlations for the b parameters were almost uniformly .971. The exception was the maximum-likelihood procedure, with a correlation of .970, a trivial difference.

Compared to correlations in Table 29, these correlations were somewhat lower. It is difficult to say whether this was due to calibration or to linking errors. Both a- and b-parameter correlations were lower in analysis of the current basic data set, however, so the drop was probably due to greater calibration error.

Table 38 presents fidelity-of-calibration data for the heterogenenous condition. Means and standard deviations of item parameters, shown in columns one and two, were essentially the same as for the homogeneous condition. Differences were due to the fact that less than half of the items used in the homogeneous condition were used here.

Parameter bias statistics, shown in columns three and four, were essentially unchanged from the homogeneous condition. Changes in biases of the a-parameter means were in the third decimal place.

Table 38. Item Parameter Error--Equivalence Methods
Heterogeneous Condition Using Systematically Sampled Examinees

	True		Bias in		Absolute	RMS	
Method	Mean	SD	Mean	SD	Error	Error	R
Bayesian							
a	1.586	.500	-.159	-.005	.377	.521	.511
b	.281	1.374	.194	.593	.576	.766	.966
Progressed Bayes							
a	1.586	.500	-.100	.018	.379	.519	.507
b	.281	1.374	.166	.512	.519	.688	.967
Max. Likelihood							
a	1.586	.500	.210	.186	.505	.676	.457
b	.281	1.374	.062	.197	.335	.423	.970
Regressed M.L.							
a	1.586	.500	.087	.122	.444	.598	.469
b	.281	1.374	.095	.305	.392	.496	.971
Robust M.L.							
a	1.586	.500	.192	.138	.473	.622	.491
b	.281	1.374	.068	.198	.334	.427	.970
Rob. Reg. M.L.							
a	1.586	.500	.106	.095	.428	.567	.505
b	.281	1.374	.094	.280	.376	.488	.968
Equivalent Tests							
a	1.586	.500	-.005	.029	.370	.507	.526
b	.281	1.374	-.016	.361	.421	.589	.956
No Linking							
a	1.586	.500	.138	.127	.455	.604	.484
b	.281	1.374	.146	.246	.368	.466	.971

Changes in the biases of the *b*-parameter means were in the second decimal place. Changes in the bias of the *a*- and *b*-parameter standard deviations were somewhat greater, but almost all were in the second decimal place.

The ranges of parameter errors shown in columns five and six were essentially unchanged from the homogeneous condition. Similarly, the linking procedures producing the least error were unchanged; the equivalent-tests method produced the least error in the *a* parameters and the maximum-likelihood procedure produced the least error in the *b* parameters.

The magnitude of the *a*-parameter error showed no apparent change from that observed in the data set containing randomly sampled examinees. The *b*-parameter error increased, however. These trends are similar to those of the homogeneous condition.

Correlations between true and estimated parameters generally showed a decrease from corresponding values in the homogeneous condition. This decrease was most pronounced for the *a* parameters. The highest *a*-parameter correlation was produced by the equivalent-tests method. This was followed by the Bayesian methods. The maximum-likelihood and no-linking methods produced the highest *b*-parameter correlations; the equivalent-tests methods produced the lowest. Where differences were trivial in the homogeneous condition, correlations ranged from .956 to .971 in the heterogeneous condition.

Characteristics of asymptotic ability estimates. Table 39 presents asymptotic ability estimate statistics for the homogeneous case of linking with systematically sampled examinees. The mean asymptotic ability was close to zero for most methods, but more different from zero than was observed with the randomly sampled examinees. The no-linking procedure produced estimates whose means were closest to zero; the equivalent-tests method produced estimates whose mean was farthest from zero. The regressed-maximum-likelihood procedure produced asymptotic estimates whose standard deviation was closest to 1.0; the Bayesian procedures produced estimates with the greatest bias in the standard deviation.

Absolute and root-mean-square errors are presented in columns three and four in Table 39. The smallest amount of error was produced by the regressed and the robust-regressed maximum-likelihood procedures; the largest error was produced by the equivalent-tests procedure. The remaining maximum-likelihood and the no-linking procedures produced errors slightly greater than the regressed and robust-regressed procedures. The Bayesian procedures produced error in an amount nearly midway between the maximum-likelihood procedures and the equivalent-tests procedure. This ordering of procedures was somewhat different from that observed in the set of randomly sampled examinees.

Table 39. Asymptotic Ability Estimates--Equivalence Methods Homogeneous Condition Using Systematically Sampled Examinees

Method	Mean	SD	Absolute Error	RMS Error	R
Bayesian	-.044	1.152	.167	.223	.996
Progressed Bayes	-.049	1.108	.145	.192	.996
Max. Likelihood	-.060	.944	.128	.176	.996
Regressed M.L.	-.054	1.003	.121	.159	.996
Robust M.L.	-.064	.936	.127	.171	.996
Rob. Reg. M.L.	-.060	.978	.121	.159	.996
Equivalent Tests	-.200	1.022	.244	.356	.996
No Linking	.003	.970	.125	.162	.996

The correlations between true and asymptotic ability were uniformly .996. This was a slight decrease from Table 31 where they were almost all .999.

Asymptotic estimate statistics for the heterogeneous condition are presented in Table 40. Slight changes from Table 39 appeared in the means, but the no-linking method still produced the least bias and the equivalent-tests method produced the most. Slight changes also occurred in the standard deviations but none were of any consequence.

In the heterogeneous condition, the no-linking procedure produced the least absolute and root-mean-square errors of the parameter estimates. The maximum-likelihood procedures were typically next in line but the Bayesian procedures closed the gap considerably. The equivalent-tests procedure still produced the most error. Root-mean-square error was invariably less for the heterogeneous condition than it had been for the homogeneous condition. Absolute error typically exhibited the same behavior but a few increases were observed. This decrease was similiar to the one observed in the data set containing randomly sampled examinees.

The correlations between true and asymptotic ability ranged from .995 to .996. These were too close in value to make any meaningful contrast between methods. The decrease from the homogeneous condition was extremely slight.

Table 40. Asymptotic Ability Estimates--Equivalence Methods
Heterogeneous Condition Using Systematically Sampled Examinees

Method	Mean	SD	Absolute Error	RMS Error	R
Bayesian	-.051	1.143	.144	.195	.996
Progressed Bayes	-.056	1.100	.121	.166	.996
Max. Likelihood	-.075	.928	.130	.157	.995
Regressed M.L.	-.066	.992	.107	.136	.996
Robust M.L.	-.076	.930	.132	.158	.995
Rob. Reg. M.L.	-.071	.972	.114	.142	.995
Equivalent Tests	-.207	1.022	.216	.231	.996
No Linking	-.013	.962	.095	.127	.995

Efficiency of ability estimation. Table 41 presents calibration and linking efficiencies for the homogeneous condition with systematically sampled examinees. The first entry in the first column indicates that slightly less information was available from true parameters in this data set than for the randomly sampled examinees (.314 vs. .319 units per item). Efficiency of calibration, as indicated by the first entry in the second column, was also slightly less (.887 vs. .898).

Linking efficiencies, presented in the third column (Table 41), were somewhat lower than those obtained with randomly sampled examinees (Table 33) and also somewhat more variable. In general, the equivalent-tests method produced the highest relative efficiency, .971. This was slightly higher than it produced in the random sampling environment. The Bayesian methods were next, both with .964. The maximum-likelihood methods ranged from .956 to .961. The no-linking procedure resulted in an efficiency of .957. By way of comparison, except for the equivalent-tests method, efficiencies in the random sampling environment were .988 to .989.

Table 42 presents relative efficiencies for the heterogeneous condition. The calibration efficiency, .889, was essentially unchanged (as it should have been since any change would be due solely to sampling). Linking efficiencies were all lower in this condition,

Table 41. Efficiency Analysis--Equivalence Methods
Homogeneous Condition Using Systematically Sampled Examinees

Method	Average Item Information	Efficiency Relative to	
		True Parameters	Estimated Parameters
True Parameters	.314		
Est. Parameters	.278	.887	
Bayesian	.268	.855	.964
Progressed Bayes	.268	.855	.964
Max. Likelihood	.267	.850	.958
Regressed M.L.	.267	.853	.961
Robust M.L.	.266	.849	.956
Rob. Reg. M.L.	.267	.851	.959
Equivalent Tests	.270	.862	.971
No Linking	.266	.849	.957

with the maximum-likelihood procedure being the lowest, .904. The equivalent-tests procedure produced the highest efficiency, .949, but the Bayesian procedure was close, .942.

All equivalent-groups and the no-linking procedures had lower efficiencies in the systematic sampling environment than in the random sampling environment. This was expected since a theoretically crucial assumption was violated. The equivalent-tests method lost no efficiency, as should also have been expected since no assumption violations occurred.

Table 43 presents linking efficiency of the Bayesian equivalent-groups method as a function of test length and sample size. Efficiencies appeared to increase with increasing sample size, but this trend was not smooth and was somewhat inconsistent when the 12 cell entries were compared. No trend with test length was obvious. Again, essentially no trends were observed in the randomly sampled data set.

Table 44 presents linking efficiency of the equivalent-tests method as a function of test length and sample size. No trend with

Table 42. Efficiency Analysis--Equivalence Methods Heterogeneous Condition Using Systematically Sampled Examinees

Method	Average Item Information	Efficiency Relative to True Parameters	Efficiency Relative to Estimated Parameters
True Parameters	.305		
Est. Parameters	.271	.889	
Bayesian	.255	.837	.942
Progressed Bayes	.255	.835	.940
Max. Likelihood	.245	.804	.904
Regressed M.L.	.249	.816	.918
Robust M.L.	.250	.819	.922
Rob. Reg. M.L.	.252	.828	.932
Equivalent Tests	.257	.844	.949
No Linking	.248	.814	.916

Table 43. Cellwise Efficiency Analysis Bayesian Score--Systematically Sampled Examinees

Sample Size	Item Set Size 20	35	50	65	Average
500	.961	.917	.954	.970	.951
1000	.969	.939	.990	.982	.970
2000	.966	.971	.994	.950	.970
Average	.965	.942	.979	.967	

Table 44. Cellwise Efficiency Analysis
Equivalent Tests--Systematically Sampled Examinees

Sample Size	Test Length				Average
	20	35	50	65	
500	.969	.907	.985	.990	.963
1000	.977	.990	.978	.992	.984
2000	.926	.957	.991	.986	.965
Average	.957	.951	.985	.989	

respect to sample size was obvious. Efficiency did appear to increase with test length in the marginal entries, although this trend was inconsistent in the individual rows. These findings regarding trends are consistent with those for the randomly sampled data set.

Discussion

Many of the data presented in this section were conflicting and inconsistent. Depending on which analyses were done, the different methods varied from best to worst. Fidelity analyses suggested that the equivalent-tests method was best and the maximum-likelihood procedure was second best. Evaluation of asymptotic ability estimates suggested that the equivalent-tests method produced the greatest asymptotic error of estimation. Efficiency analyses suggested that the equivalent-tests method was most efficient and the Bayesian procedures were almost as efficient.

The efficiency analysis probably produces the best answers to questions of which procedure is best. It is the goal of linking, after all, to produce a set of items that will function efficiently together. The facts that the parameters are not "most true" or that the ability scale is not at arbitarily targeted levels are secondary to the goal of efficiency of measurement. Efficiency analyses are probably most useful in selecting a procedure.

Accepting the previous argument, several observations can be made. First, the equivalent-tests method is the most efficient when examinees are systematically sampled, as they were here. Second, the Bayesian procedures are nearly as efficient with systematic sampling and, as was observed earlier, are more efficient when examinees are randomly sampled. At some point between the extremes in sampling investigated here, the Bayesian procedures could be expected to become superior. Of the two Bayesian procedures, neither was clearly superior, but the simple (i.e., unprogressed) procedure was easier to compute and therefore preferable.

Analysis of the two methods by test length and sample size suggested that there was a slight increase in efficiency of the equivalent-tests method as test length increased and a slight increase in efficiency of the Bayesian equivalent-groups procedure as sample size increased. These increases were small and inconsistent, however, and suggested that all of the test lengths and sample sizes investigated were nearly equivalent in terms of resulting efficiency for both the equivalent-tests and Bayesian methods.

Anchor Group Method

Procedure

The anchor group linking method is, conceptually, very similar to the equivalent-groups method. The major conceptual distinction is that the anchor group method uses a single group of examinees for all linking and thus does not need to assume the statistical equivalence of several different groups.

In this research, eight different anchor groups were evaluated. The eight groups comprised four examinee sample sizes (10, 30, 50, and 100) and two distribution forms (rectangular and normal). The rectangular samples consisted of abilities evenly spaced between -1.7 and 1.7. The normal samples were created by selecting normal deviates corresponding to evenly spaced percentiles from 2.0 to 98.0. Values thus obtained for both normal and rectangular samples were then standardized to assure that the samples obtained had means of exactly zero and variances of exactly one.

Linking by the anchor group method was done for all parameters in the systematically sampled data set. This was accomplished by administering all 60 tests in the data set to each of the examinees in each of the anchor groups. Item parameters were then adjusted using the same equations used for the equivalent groups method, Equations 14 and 15. Two scoring procedures, the modal Bayesian procedure and the robust-maximum-likelihood procedure were used for this linking.

Results--Modal Bayesian Scores

Fidelity of parameter estimation. Table 45 presents the item parameter error statistics for the anchor group linking method for each anchor group size and composition in the homogeneous linking condition using modal Bayesian estimates. The first two columns present the means and standard deviations of the true a and b parameters averaged over cells in the systematically sampled data set. These

Table 45. Item Parameter Error--Anchor Groups
Homogeneous Condition Using Systematically Sampled Examinees

	True		Bias in		Absolute	RMS	
Method	Mean	SD	Mean	SD	Error	Error	R
Normal 10							
a	1.588	.501	-.080	.033	.393	.540	.519
b	.262	1.344	.180	.479	.440	.671	.977
Normal 30							
a	1.588	.501	-.076	.017	.380	.521	.527
b	.262	1.344	.168	.443	.409	.614	.979
Normal 50							
a	1.588	.501	-.086	.019	.381	.525	.529
b	.262	1.344	.186	.469	.424	.644	.979
Normal 100							
a	1.588	.501	-.101	.011	.374	.516	.530
b	.262	1.344	.193	.480	.432	.659	.979
Uniform 10							
a	1.588	.501	-.110	.024	.395	.545	.516
b	.262	1.344	.198	.516	.470	.717	.976
Uniform 30							
a	1.588	.501	-.135	.006	.386	.529	.520
b	.262	1.344	.192	.530	.469	.706	.977
Uniform 50							
a	1.588	.501	-.137	.001	.378	.523	.529
b	.262	1.344	.203	.530	.470	.712	.979
Uniform 100							
a	1.588	.501	-.115	.003	.372	.516	.531
b	.262	1.344	.208	.497	.448	.681	.980
No Linking							
a	1.588	.501	.139	.084	.450	.602	.533
b	.262	1.344	.130	.237	.364	.464	.971

values are the same as those presented in Table 37 and will not be discussed again here.

Biases in the estimated item parameters are presented in columns three and four. With the exception of the no-linking group, all groups tended to underestimate the *a* parameters. All groups tended to overestimate the *b* parameters, with a trend for increasing bias with increasing group size. The no-linking method revealed the least *b*-parameter bias, while the normal group showed the least bias overall. In terms of bias in parameter standard deviations, the uniform group showed least bias in the *a* parameters and the normal group showed least bias in the *b* parameters. Again, the no-linking method showed the least bias in the *b* parameters overall.

Absolute and root-mean-square errors of the parameter estimates are presented in columns five and six. A slight trend toward decreasing absolute error in the *a* parameters with increasing anchor group size was apparent for both distributions, although it was more pronounced with the uniform anchor groups. No consistent differences were apparent between the group compositions with respect to *a*-parameter absolute error, but both produced less error than the no-linking procedure. Absolute error of the *b* parameters suggested different conclusions: There were no noticeable decreases with increasing anchor group sizes for the normal group and there were slight decreases for the uniform group. The no-linking procedure produced the least error, and the uniform groups consistently produced the most error. The same conclusions drawn from the absolute errors could also be drawn from the root-mean-square errors.

The correlations between true and estimated *a* and *b* parameters are shown in the last column of Table 45. There was a slight increasing trend in both the *a*- and *b*-parameter correlations with increasing anchor group size for both shapes of ability distribution. The no-linking procedure produced *a*-parameter correlations slightly higher than those of other methods and *b*-parameter correlations that were slightly lower.

The fidelity-of-calibration data for the heterogeneous condition are presented in Table 46. Since observations about the true item parameters remain the same across linking methods, they will not be repeated here.

The parameter biases presented in columns three and four were essentially the same as those of the homogeneous case. The bias of the *a*-parameter means tended to be somewhat smaller for the homogeneous case while the same trend was observed with respect to bias in the a-parameter standard deviations. For the *b* parameters, however, the bias in both the mean and standard deviation were greater in the heterogeneous condition.

Parameter errors depicted in columns five and six were essentially the same as those for the homogeneous case for the *a* parameters. The differences between the heterogeneous and homogeneous conditions appeared in the third decimal place for the *a* parameters. The *b*-parameter errors for the heterogeneous condition showed a

Table 46. Item Parameter Error--Anchor Groups Heterogeneous Condition Using Systematically Sampled Examinees

	True		Bias in		Absolute	RMS	
Method	Mean	SD	Mean	SD	Error	Error	R
Normal 10							
a	1.586	.500	-.082	.045	.394	.538	.497
b	.281	1.374	.203	.501	.450	.680	.972
Normal 30							
a	1.586	.500	-.077	.027	.384	.522	.507
b	.281	1.374	.189	.468	.419	.622	.974
Normal 50							
a	1.586	.500	-.087	.029	.385	.526	.501
b	.281	1.374	.207	.492	.435	.653	.974
Normal 100							
a	1.586	.500	-.102	.017	.377	.517	.515
b	.281	1.374	.214	.504	.443	.667	.973
Uniform 10							
a	1.586	.500	-.111	.040	.400	.547	.477
b	.281	1.374	.219	.550	.483	.730	.968
Uniform 30							
a	1.586	.500	-.137	.011	.389	.530	.498
b	.281	1.374	.215	.557	.482	.718	.972
Uniform 50							
a	1.586	.500	-.138	.006	.381	.524	.505
b	.281	1.374	.224	.557	.482	.721	.972
Uniform 100							
a	1.586	.500	-.117	.008	.374	.516	.513
b	.281	1.374	.229	.525	.459	.690	.973
No Linking							
a	1.586	.500	.138	.127	.455	.604	.484
b	.281	1.374	.146	.246	.368	.466	.971

slight increase over the homogeneous condition. Absolute errors of the b parameters showed no noticeable trends with increasing anchor group size for the normal groups but showed a slight decreasing trend with increasing uniform anchor group size. Many of the same conclusions could also be drawn from the root-mean-square errors.

Whereas bias and error statistics were quite similar for the homogeneous and heterogeneous conditions, the correlations between true and estimated parameters showed a noticeable drop from their corresponding values in the homogeneous condition. Differences in the second decimal place were observed for the *a* parameters and in the third decimal place for the *b* parameters. There was a slight tendency for the correlations to increase with increasing anchor group size. The no-linking procedure's correlation for the *a* parameters was, however, somewhat lower than most correlations produced by the anchor group procedures.

Characteristics of asymptotic ability estimates. Table 47 presents descriptive statistics for the asymptotic ability estimates in the homogeneous case. Mean asymptotic ability estimates were close to zero for all cases while the corresponding standard deviations were close to one. For the most part, means were overestimated, as were the standard deviations.

Table 47. Asymptotic Ability Estimates--Anchor Groups
Homogeneous Condition Using Systematically Sampled Examinees

Method	Mean	SD	Absolute Error	RMS Error	R
Normal 10	.005	1.070	.085	.131	.996
Normal 30	-.009	1.066	.081	.129	.996
Normal 50	.004	1.070	.081	.129	.996
Normal 100	.004	1.078	.081	.134	.995
Uniform 10	.003	1.092	.105	.156	.996
Uniform 30	-.005	1.104	.101	.151	.996
Uniform 50	.005	1.108	.098	.157	.996
Uniform 100	.017	1.091	.085	.142	.996
No Linking	.003	.970	.125	.162	.996

Absolute error presented in column three was lowest for the normal anchor group and greatest for the no-linking procedure. Absolute error appeared to decrease with increasing anchor group size for the uniform anchor group. No trend was obvious for the normal group.

Root-mean-square error, presented in column four, showed the same differences among linking methods. Trends within methods as a function of anchor group size were not apparent.

Correlations between the true and asymptotic ability, shown in column five, were uniformly .996.

Statistics for the asymptotic ability in the heterogeneous case are presented in Table 48. Slight changes were observed from the homogeneous condition, for the means and standard deviations. Whereas the homogeneous condition tended to overestimate the means, the heterogeneous condition tended to underestimate them. Standard deviations of the asymptotic estimates for the heterogeneous condition were smaller than for the homogeneous condition.

Table 48. Asymptotic Ability Estimates--Anchor Groups
Heterogeneous Condition Using Systematically Sampled Examinees

Method	Mean	SD	Absolute Error	RMS Error	R
Normal 10	.004	1.065	.085	.125	.996
Normal 30	-.012	1.061	.075	.117	.996
Normal 50	-.001	1.066	.072	.117	.996
Normal 100	.000	1.075	.078	.125	.996
Uniform 10	-.000	1.082	.085	.130	.996
Uniform 30	-.009	1.100	.096	.139	.996
Uniform 50	-.001	1.103	.095	.140	.996
Uniform 100	.014	1.088	.081	.131	.996
No Linking	-.013	.962	.095	.127	.995

Absolute and root-mean-square errors of the asymptotic estimates were uniformly lower in the heterogeneous condition than in the homogeneous condition. Trends with respect to anchor group size were not apparent, however, and the no-linking method was not consistently inferior.

Correlations between true and asymptotic ability were identical to the homogeneous condition (i.e., .996) for the anchor group procedures. The no-linking procedure produced a correlation slightly lower in the heterogeneous condition.

Efficiency of ability estimation. Table 49 presents the efficiencies achieved by the homogeneous linking condition with systematically sampled examinees. The average item information, presented in the first column, was nearly identical for both the normal and uniform groups and increased as sample size increased. The no-linking group showed the lowest average item information.

Table 49. Efficiency Analysis--Anchor Groups
Homogeneous Condition Using Systematically Sampled Examinees

Method	Average Item Information	Efficiency Relative to True Parameters	Efficiency Relative to Estimated Parameters
True Parameters	.314		
Est. Parameters	.278	.887	
Normal 10	.272	.869	.979
Normal 30	.274	.875	.986
Normal 50	.274	.875	.986
Normal 100	.275	.876	.987
Uniform 10	.272	.866	.976
Uniform 30	.274	.873	.983
Uniform 50	.275	.876	.987
Uniform 100	.275	.877	.988
No Linking	.266	.849	.957

Linking efficiency, shown in the third column, showed a slight rise as sample size went from 10 to 30 but negligible change from 30 to 100. There were no consistent differences between the two anchor group distributions. The no-linking case showed the lowest efficiency, .957.

Relative efficiencies for the heterogeneous condition are presented in Table 50. The same trends were apparent here (except for rounding error) as were shown for the homogeneous case. Information values and relative efficiencies were markedly lower for the heterogeneous condition than for the homogeneous condition. As before, a sharp rise was noted as sample size increased from 10 to 30, but there were negligible increases thereafter.

Table 50. Efficiency Analysis--Anchor Groups
Heterogeneous Condition Using Systematically Sampled Examinees

Method	Average Item Information	Efficiency Relative to	
		True Parameters	Estimated Parameters
True Parameters	.305		
Est. Parameters	.271	.889	
Normal 10	.259	.850	.956
Normal 30	.261	.857	.964
Normal 50	.260	.855	.962
Normal 100	.261	.858	.966
Uniform 10	.257	.845	.951
Uniform 30	.261	.856	.963
Uniform 50	.261	.858	.966
Uniform 100	.262	.860	.968
No Linking	.248	.814	.916

Results--Robust-Maximum-Likelihood Scores

Fidelity of parameter estimation. Table 51 is a condensed table of the modal Bayesian and robust-maximum-likelihood item parameter error statistics for the anchor group linking design in the homogeneous

Table 51. Item Parameter Error--Anchor Groups
Homogeneous Condition Using Systematically Sampled Examinees

	Bayesian				Maximum Likelihood			
	Bias in		RMS		Bias in		RMS	
Method	Mean	SD	Error	R	Mean	SD	Error	R
Normal 10								
a	-.054	.060	.562	.489	.699	.391	1.168	.438
b	.151	.422	.590	.978	-.035	-.118	.331	.973
Normal 30								
a	-.076	.037	.552	.488	.454	.256	.834	.444
b	.164	.429	.597	.981	-.004	.007	.426	.968
Normal 50								
a	-.052	.051	.562	.486	.441	.244	.857	.467
b	.166	.419	.597	.979	-.016	.002	.320	.975
Normal 100								
a	-.107	.025	.541	.487	.483	.263	.896	.462
b	.203	.468	.653	.980	-.023	-.027	.307	.976
Uniform 10								
a	-.060	.066	.601	.463	-.007	.182	.706	.381
b	.185	.447	.637	.975	.160	.531	.905	.952
Uniform 30								
a	-.127	.023	.549	.483	.120	.165	.640	.478
b	.182	.500	.671	.979	.071	.300	.581	.971
Uniform 50								
a	-.117	.030	.555	.485	.175	.174	.717	.457
b	.207	.499	.684	.979	.079	.222	.426	.974
Uniform 100								
a	-.105	.028	.546	.487	.169	.160	.670	.453
b	.207	.478	.673	.980	.072	.232	.497	.973
No Linking								
a	.143	.112	.629	.501	.143	.112	.629	.501
b	.147	.228	.444	.973	.147	.228	.444	.973

case. The table values represent averages taken over four cells of the data matrix (i.e. 1000 examinees and 20, 35, 50, and 65 items), rather than over the entire 3x4 matrix, as in the previous section.

Whereas the bias in the a-parameter means, using modal Bayesian estimation, tended to be slightly negative for both the normal and uniform groups (indicating that the a parameters were underestimated), the robust-maximum-likelihood procedure grossly overestimated the means for the normal group and slightly overestimated the means for the uniform group. The trends with respect to the b-parameter biases were reversed from those noted for the a parameters. The robust-maximum-likelihood procedure produced a b-parameter mean that was much closer to the true value of 0.0 than did the modal Bayesian estimate. The normal group tended to produce slight underestimates of the b-parameter mean while the uniform group produced slight overestimates. Both groups produced overestimates of the b mean when the modal Bayesian scoring procedure was used.

The same general trends noted for the bias in parameter means held also for the biases in the parameter standard deviations. The robust-maximum-likelihood estimates tended to overestimate the a-parameter standard deviations more than their counterparts in the Bayesian case. As was the case for the b-parameter means, the robust-maximum-likelihood estimates of the standard deviations were much closer to the true value of 1.0 than were the modal Bayesian estimates. The normal groups revealed a much smaller bias in b-parameter standard deviations than did the uniform groups using robust maximum likelihood. The Bayesian modal estimates showed very little difference between the normal and uniform groups.

In terms of root-mean-square error in the a parameter, modal Bayesian procedures showed the least error, regardless of distribution shape. On the other hand, robust-maximum-likelihood procedures provided the smallest errors for the b parameters. The normal group produced less error than the uniform group, with a slight tendency for increasing error with increasing anchor group size.

The correlations between true and estimated parameters were consistently higher with modal Bayesian procedures than with robust-maximum-likelihood procedures although in several instances the differences were in the third decimal place. There were no consistent differences among group compositions or sizes. As usual, correlations for the b parameters were considerably higher than for the a parameters.

Characteristics of asymptotic ability estimates. Table 52 presents summary statistics for the asymptotic ability estimates using both modal Bayesian and robust-maximum-likelihood procedures. The robust-maximum-likelihood procedure resulted in slight underestimation of the means for both the normal and uniform groups. Standard deviations were also underestimated, compared to the modal Bayesian groups which tended to overestimate the standard deviation. For the robust-maximum-likelihood procedures, there was a noticeable difference between the normal group, which produced underestimated standard

Table 52. Asymptotic Ability Estimates--Anchor Groups Homogeneous Condition Using Systematically Sampled Examinees

Method	Bayesian Mean	Bayesian SD	Bayesian RMS Error	Bayesian R	Maximum Likelihood Mean	Maximum Likelihood SD	Maximum Likelihood RMS Error	Maximum Likelihood R
Normal 10	-.006	1.045	.114	.996	-.037	.724	.305	.996
Normal 30	-.009	1.066	.125	.996	-.068	.776	.258	.996
Normal 50	-.004	1.044	.108	.996	-.048	.791	.236	.996
Normal 100	.010	1.080	.131	.996	-.044	.779	.247	.996
Uniform 10	.013	1.061	.126	.997	-.048	.993	.136	.997
Uniform 30	-.009	1.098	.144	.996	-.049	.932	.133	.997
Uniform 50	.012	1.090	.134	.996	-.015	.920	.143	.996
Uniform 100	.016	1.078	.128	.996	-.033	.911	.135	.996
No Linking	.034	.962	.133	.996	.034	.962	.133	.996

deviations, and the uniform group, which produced overestimated standard deviations.

In terms of root-mean-square error, there were again notable differences between the normal and uniform groups using robust-maximum-likelihood procedures. The normal group had bias values considerably greater than its counterpart using modal Bayesian procedures while the uniform group had error values quite comparable to their Bayesian counterparts. The normal-group errors, using robust-maximum-likelihood scoring, were by far the largest of any of the methods.

Correlations between true and estimated parameters using robust-maximum-likelihood procedures were uniformly high (.996) and virtually identical to their Bayesian counterparts.

Efficiency of ability estimation. Table 53 presents comparisons of robust-maximum-likelihood with modal Bayesian procedures in terms of relative efficiencies achieved by each method. The average amount of information available per item tended to be higher for the modal Bayesian procedures than for the robust-maximum-likelihood procedures. This, of course, meant that the efficiencies relative to the true and

Table 53. Efficiency Analysis--Anchor Groups
Homogeneous Condition Using Systematically Sampled Examinees

	Bayesian			Maximum Likelihood		
		Efficiencies Relative to			Efficiencies Relative to	
Method	Avg. Item Info.	True Params.	Est. Params.	Avg. Item Info.	True Params.	Est. Params.
True Params.	.306			.306		
Est. Params.	.270	.882		.270	.882	
Normal 10	.265	.866	.983	.257	.840	.953
Normal 30	.267	.874	.991	.262	.857	.972
Normal 50	.266	.870	.987	.265	.868	.984
Normal 100	.267	.873	.991	.264	.862	.978
Uniform 10	.263	.860	.976	.252	.824	.935
Uniform 30	.267	.872	.989	.262	.856	.971
Uniform 50	.267	.872	.990	.262	.858	.973
Uniform 100	.267	.873	.990	.264	.865	.981
No Linking	.260	.850	.964	.260	.850	.964

estimated parameters were also higher for modal Bayesian than for robust-maximum-likelihood procedures. The magnitude of differences were, with one exception, in the second decimal place.

The normal group showed no consistent trend with increasing group size. The uniform group showed a tendency for increasing efficiency with increasing group size. These trends appeared for both modal Bayesian and robust-maximum-likelihood procedures.

Discussion

Most of the analyses thus far have presented rather conflicting results. Different analyses have suggested different procedures that were "best." Using fidelity-of-parameter estimation as a criterion, modal Bayesian procedures tended to produce more accurate estimates of the _a_ parameter while the robust-maximum-likelihood procedures tended to produce more accurate estimates of the b parameter. Within the modal Bayesian procedures, there did not appear to be any clear-cut advantage to either group composition. For the robust-maximum-likelihood procedures, there was a clear trend for the normal groups to produce consistently better estimates for the b parameters than those estimates produced from the uniform groups.

Using asymptotic ability estimates as the evaluative criterion, modal Bayesian procedures with normally distributed anchor group abilities appeared to be consistently best. Modal Bayesian procedures with uniformly distributed abilities were second best. Robust-maximum-likelihood scoring using uniform and normal anchor groups followed in that order.

Modal Bayesian procedures showed efficiencies consistently higher than robust-maximum-likelihood procedures regardless of anchor group composition or size. With the modal Bayesian procedures, the normal groups tended to yield slightly more efficiency than did the uniform groups. Both groups were superior to the no-linking condition.

Anchor Test Method

Procedure

Generation of the source item pool. The first step in the application of the anchor test method was to construct a source item pool from which the anchor tests could be selected. To obtain the source item pool, 200 a, b, and c parameters were independently generated as discussed previously. The first four central moments of each of these distributions matched those specified earlier as being representative of a "typical" ASVAB item pool. These parameters represented the "true" parameters of 200 hypothetical items.

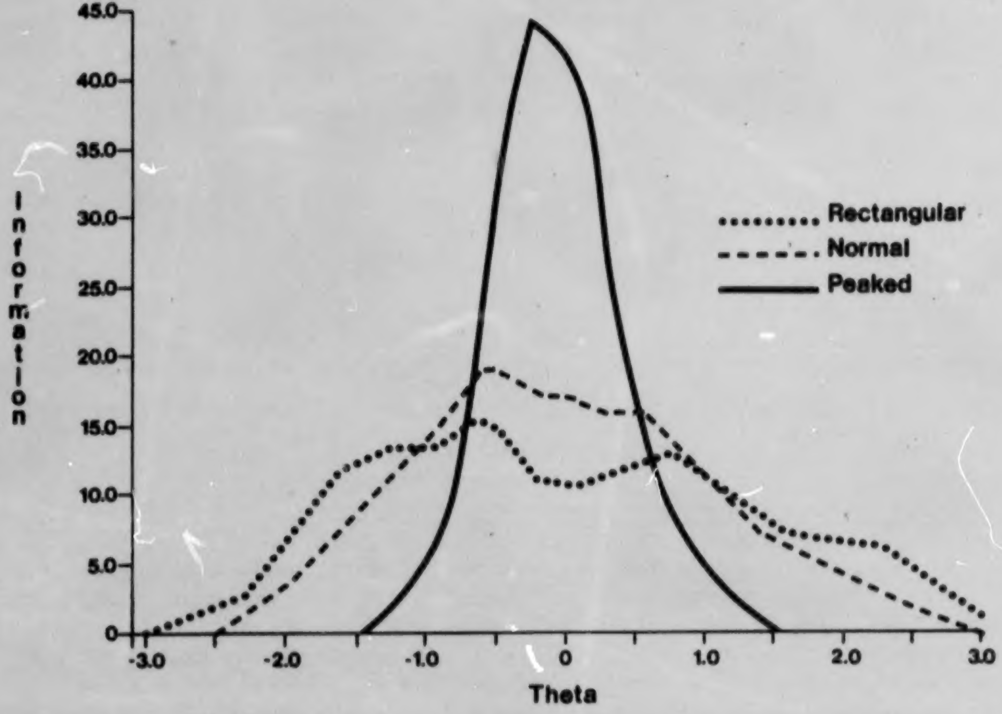
Dichotomous item responses for these 200 items were simulated for 4000 examinees randomly selected from a distribution of abilities with distributional moments representative of the total AFEES population. All examinees responded according to the three-parameter logistic IRT model. Item parameter estimates were obtained for these 200 items using program OGIVIA. The items were, due to computer program limitations, calibrated in two sets of 100 items each.

Selection of anchor-test items. Three different 25-item anchor tests were constructed by selecting items from the original set of 200 items. These anchor tests were constructed so that their test information curves were approximately normal, rectangular, and peaked.

The peaked test was constructed by selecting the 25 items which provided the most information at theta equal to zero, according to their estimated item parameters; this is the way items would typically be selected for inclusion in a peaked test. In order to get an indication of the amount of information actually contained in this test, the true information was computed, using the true item parameters, for 61 theta values at intervals of .10 from -3.00 to 3.00. These information values were then averaged across 61 theta values; this average was 8.320.

Items for the rectangular and normal tests were selected so that their test information curves were shaped approximately rectangular and normal, respectively, and so that the true test information, computed using the true item parameters and averaged as before over 61 theta values from -3.00 to 3.00, approached the value obtained by the peaked test. These averages were 8.410 and 8.232 for the rectangular and normal tests, respectively. When the test information was computed on the basis of the estimated item parameters, these averages were 8.485, 9.294, and 9.121 for the peaked, rectangular, and normal tests, respectively. Figure 9 presents the true information curves, based on the true item parameters, for the three 25-item anchor tests.

Figure 9. True Information Curves, Using True Item Parameters, for Each of Three Anchor Tests



Two additional embedded tests for each of these three anchor tests were obtained by selecting the first five items and the first 15 items from each. Thus, the nine anchor tests considered here comprised three groups of 5-, 15-, and 25-item tests, each of whose test information curves for these tests were approximately normal, rectangular, and peaked, respectively. The items included in these anchor tests are presented in Appendix Table A-2.

Determination of the linking transformations. The nine anchor tests were "administered" to the 70,000 examinees comprising the systematically sampled basic data set. This simulation was accomplished by generating response vectors using the true theta levels of these examinees and then scoring the anchor tests. Once item responses were available for the items in each anchor test, a modal Bayesian estimate of ability was computed for each examinee on each anchor test, using a standard normal prior distribution of abilities and scoring each response vector using the estimated item parameters. For each of the 60 calibration groups, the mean and standard deviation of estimated ability were computed on each of the nine anchor tests. These values were then used for the transformation constants for anchor-test linking.

Linking under the anchor-test method is accomplished by transforming the non-anchor-test item parameters such that the mean and standard deviation of ability of the groups under consideration, as estimated from the non-anchor test, match the mean and standard deviation of ability estimated from the anchor test alone. When the transformation constants k and m are applied in the form presented by Equations 14 and 15, the constants k and m may be expressed as:

$$k = \sigma_\Gamma / \sigma_\theta \qquad [30]$$

and

$$m = \mu_\Gamma - k\mu_\theta \qquad [31]$$

where μ_Γ and σ_Γ are, respectively, the mean and standard deviation of ability estimates in the non-anchor test and μ_θ and σ_θ are the corresponding statistics for the anchor test.

Results--Modal Bayesian Scores

Fidelity of parameter estimation. Fidelity-of-estimation statistics for the homogeneous condition, using the Bayesian scoring technique, are presented in Table 54. The true means and standard deviations of the a and b parameters are presented in the first two columns of this table. Columns three and four present the bias in the means and standard deviations of the item parameters. The largest

Table 54. Item Parameter Error--Anchor Tests
Homogeneous Condition Using Systematically Sampled Examinees

Method	True		Bias in		Absolute	RMS	
	Mean	SD	Mean	SD	Error	Error	R
Normal 5							
a	1.588	.501	.574	.237	.718	.874	.532
b	.262	1.344	.135	-.091	.258	.350	.979
Normal 15							
a	1.588	.501	.095	.076	.414	.552	.531
b	.262	1.344	.226	.266	.320	.509	.980
Normal 25							
a	1.588	.501	.067	.067	.405	.544	.530
b	.262	1.344	.232	.293	.333	.529	.980
Rectangular 5							
a	1.588	.501	.400	.182	.589	.738	.530
b	.262	1.344	.168	.020	.253	.365	.980
Rectangular 15							
a	1.588	.501	.095	.077	.416	.554	.532
b	.262	1.344	.227	.267	.321	.506	.980
Rectangular 25							
a	1.588	.501	.042	.058	.396	.536	.531
b	.262	1.344	.233	.318	.344	.544	.980
Peaked 5							
a	1.588	.501	1.092	.418	1.169	1.359	.531
b	.262	1.344	.029	-.332	.342	.430	.980
Peaked 15							
a	1.588	.501	.617	.255	.754	.914	.531
b	.262	1.344	.102	-.115	.255	.344	.980
Peaked 25							
a	1.588	.501	.457	.201	.629	.780	.529
b	.262	1.344	.145	-.017	.248	.359	.979
No Linking							
a	1.588	.501	.139	.084	.450	.602	.533
b	.262	1.344	.130	.237	.364	.464	.971

biases in the mean of the a parameters were observed for the peaked tests, and ranged from .457 for the 25-item anchor test to 1.092 for the 5-item anchor test. The smallest biases in the means were observed for the rectangular tests, although the biases for the normal tests were only slightly higher at the longer test lengths. The smallest biases were observed for the 25-item normal and rectangular tests, with values of .067 and .042, respectively. When no linking was performed, the bias in the mean of the a parameters was .139; this value was exceeded by all three peaked tests, but only by the 5-item normal and rectangular tests.

Biases in the standard deviations of the a parameters were largest for the peaked tests, ranging from .201 to .418. Again, there was little difference observed between the biases in the standard deviations of the a parameters for the normal and the rectangular tests, although they were slightly smaller for the rectangular tests. The smallest biases were observed for the 25-item normal and rectangular tests. As before, biases for all three peaked tests exceeded the value of .084 observed in the no-linking condition, whereas only the 5-item normal and rectangular tests exceeded this value. Biases in both the means and the standard deviations of the a parameters decreased with increased test length.

The smallest biases in the mean of the b parameters were observed for the peaked tests; these values ranged from .029 to .145. There were essentially no differences between the rectangular and normal tests in terms of bias in the mean b's; these values clustered between .135 and .233. These bias figures increased with increased test lengths for all three anchor test types. In the no-linking condition, bias in the mean b's was .130, which was exceeded by all tests except the 5- and 15-item peaked tests.

The standard deviations of the b parameters were underestimated for the peaked tests, since all these bias values were negative, ranging from -.017 to -.332. The differences between the normal and rectangular tests were not consistent, though the normal test was somewhat better at test lengths greater than five items. The bias in the b-parameter standard deviation was .237 in the no-linking condition, and this value was exceeded by all the tests except the shortest normal and rectangular tests and the two longest peaked tests.

Mean absolute and root-mean-square errors in the parameters are presented in columns five and six of Table 54. The peaked anchor tests performed most poorly according to both of these indices of error for the a parameters. The mean absolute error in estimating a was .629 for the 25-item peaked test, and was as high as 1.169 for the 5-item peaked test. The rectangular tests were best overall, but for 15 and 25 items, the normal tests performed nearly as well. The least error was observed for the 25-item rectangular and normal tests. When no linking was performed at all, mean absolute error was .450.

All three peaked tests exceeded this value, but only the 5-item version of the normal and rectangular tests did.

The pattern was identical for the root-mean-square error in the a parameters. That is, the peaked tests performed most poorly, and all three peaked tests exceeded the root-mean-square error of .602 which was observed in the no-linking condition. Again, the rectangular tests were best overall, but for 15 and 25 items, the normal tests performed nearly as well. The least error was observed for the 25-item rectangular and normal tests. For all three kinds of anchor tests, both absolute and root-mean-square errors in the a parameters decreased with increasing anchor test size.

The pattern of errors was somewhat different for the b parameters. Overall, there were essentially no differences among the anchor test types in mean absolute error; these values ranged from .248 to .344 across the nine tests, and all these values were below the .364 observed in the no-linking condition. For the peaked tests, mean absolute errors decreased with anchor test size as expected. For the rectangular and normal tests, however, these errors increased with test size, as was observed for the bias statistics.

The peaked tests were better, in general, than the other two kinds of tests in terms of root-mean-square errors in the b parameters. These values ranged from .344 to .430 and, although there was no trend observed with respect to anchor test size, all these values were below the .464 observed in the no-linking condition. The normal tests were slightly superior to the rectangular tests in terms of root-mean-square error. In both cases, errors increased with increasing anchor test length.

There were small differences observed across anchor tests in terms of the correlations between the true and estimated item parameters. For the a parameters, these values clustered between .529 and .532 for all nine anchor tests; all these correlations were lower than the .533 observed in the no-linking condition. There were no systematic trends observed with anchor test size.

For the b parameters, these correlations were approximately .980 for all nine tests, and therefore, all of them were higher than the .971 observed in the no-linking condition.

Fidelity-of-estimation statistics for the heterogeneous condition are presented in Table 55. As was observed for the homogeneous condition, bias in the mean a parameters was largest for the peaked tests and smallest for the rectangular tests; bias for the normal tests was only slightly larger than that for the rectangular tests. In the no-linking condition, bias in the mean a parameter was .138, which was exceeded by all the peaked tests and by the 5-item normal

Table 55. Item Parameter Error--Anchor Tests
Heterogeneous Condition Using Systematically Sampled Examinees

	True		Bias in		Absolute	RMS	
Method	Mean	SD	Mean	SD	Error	Error	R
Normal 5							
a	1.586	.500	.571	.246	.714	.871	.513
b	.281	1.374	.143	-.084	.261	.347	.975
Normal 15							
a	1.586	.500	.093	.082	.417	.552	.515
b	.281	1.374	.242	.285	.328	.514	.974
Normal 25							
a	1.586	.500	.066	.075	.410	.544	.513
b	.281	1.374	.248	.313	.341	.535	.974
Rectangular 5							
a	1.586	.500	.397	.193	.590	.738	.512
b	.281	1.374	.178	.029	.257	.363	.975
Rectangular 15							
a	1.586	.500	.093	.085	.419	.554	.515
b	.281	1.374	.242	.284	.328	.511	.975
Rectangular 25							
a	1.586	.500	.041	.065	.401	.536	.514
b	.281	1.374	.250	.338	.352	.550	.975
Peaked 5							
a	1.586	.500	1.088	.431	1.161	1.355	.512
b	.281	1.374	.032	-.332	.347	.431	.974
Peaked 15							
a	1.586	.500	.615	.266	.750	.913	.513
b	.281	1.374	.110	-.107	.258	.341	.974
Peaked 25							
a	1.586	.500	.455	.212	.628	.780	.511
b	.281	1.374	.155	-.005	.251	.358	.973
No Linking							
a	1.586	.500	.138	.127	.455	.604	.484
b	.281	1.374	.146	.246	.368	.466	.971

and rectangular tests. These bias figures decreased with increased test length for all three anchor test types.

Bias in the standard deviation of the a parameters was greatest for the peaked tests, ranging from .212 to .431. There were only small differences between the normal and rectangular tests, with the slight advantage going to the rectangular test at the longer test lengths. The smallest biases were observed for the 25-item normal and rectangular tests. The bias in the no-linking condition, .127, was exceeded by all the peaked tests and the 5-item normal and rectangular tests. As before, all these bias figures decreased with increased test lengths.

In terms of the bias in the mean b parameters, the peaked tests performed best, with bias equal to .032 for the 5-item test and increasing to .155 for the 25-item test. Bias in the mean b's was somewhat larger for the other two types of anchor tests, although there were fewer differences between them. For the normal and rectangular tests, the bias figures fell between .143 and .250. All but one of these values were greater than the .146 observed in the no-linking condition. Only the 25-item peaked test exceeded this value.

The standard deviations of the b parameters were consistently underestimated by the peaked tests; bias was as high as -.332 for the 5-item test, but was only -.005 for the 25-item test. Bias values for the other two types of tests were essentially the same, with a slight advantage going to the normal test at the longer test lengths. In the no-linking condition, bias in the standard deviation of the b parameters was .246, which was exceeded by all but the shortest normal and rectangular tests and the two longest peaked tests.

The patterns of mean absolute and root-mean-square errors in the a and b parameters in the heterogeneous condition were identical to what was observed in the homogeneous condition. In terms of mean absolute error, the peaked anchor tests performed most poorly, with errors ranging from .628 to 1.161 for the a parameter. Again, the rectangular tests were best overall, with the normal tests closely following. When no linking was performed at all, mean absolute error for the a parameter was .455. All three peaked test exceeded this value, but only the 5-item normal and rectangular tests did. This pattern of the absolute errors was repeated for the root-mean-square errors.

The pattern of errors in the b parameters for the heterogeneous case paralleled that observed in the b parameters for the homogeneous case. Overall, there were essentially no differences among the anchor test types in mean absolute error; all values were below the .368 observed in the no-linking condition. For the peaked tests, mean absolute errors decreased with anchor test size as expected. For the rectangular and normal tests, however, these errors increased with test size, as was observed for the bias statistics.

The peaked tests were better, in general, than the other two kinds of tests in terms of root-mean-square error for the b parameters. These values ranged from .341 to .431 and, although there was no trend observed with respect to anchor test size, all these values were below the .466 observed in the no-linking condition. The normal tests were slightly superior to the rectangular tests in terms of root-mean-square error. In both cases, errors increased with increased test length.

Small differences were observed across anchor tests in terms of the correlations between the true and estimated item parameters. For the a parameters, these values clustered between .511 and .515, with the lowest correlations observed for the peaked tests. All these correlations were higher than the .484 observed in the no-linking condition. There were no systematic trends observed with anchor test size.

For the b parameters, these correlations were between .973 and .975, with the lowest correlations again observed for the peaked tests. All these correlations were higher than the .971 observed in the no-linking condition.

Characteristics of asymptotic ability estimates. Table 56 presents the summary characteristics of asymptotic ability estimates for the homogeneous case. Columns 1 and 2 present the mean and standard deviation of the asymptotic ability metric. The peaked tests came closest to producing an ability metric with a mean of zero; this value increased with increased test lengths. There were essentially no differences observed between the normal and rectangular tests. For the normal tests, the means also increased with increased test length; for the rectangular tests, the means decreased.

The peaked tests performed most poorly in producing ability estimates with a standard deviation of 1.0. The rectangular tests produced estimates with a standard deviation closest to 1.0. For all three types of anchor tests, the standard deviation increased with increased test length.

The no-linking condition produced estimates whose mean, .003, was closer to zero than were the means from any of the nine anchor tests. The standard deviation for the no-linking condition, .970, was exceeded only by the 25-item normal and rectangular tests.

Although the estimates from the peaked tests had means closer to zero than did the other anchor tests, the peaked test estimates had the highest mean absolute errors. The rectangular tests had the smallest errors, but the errors for the normal tests were only slightly larger. Errors for all three peaked tests exceeded the value of .125 observed in the no-linking condition. Only the 5-item normal and rectangular tests exceeded this value. In all cases, mean absolute error decreased with increased test length. The pattern for the

Table 56. Asymptotic Ability Estimates--Anchor Tests
Homogeneous Condition Using Systematically Sampled Examinees

Method	Mean	SD	Absolute Error	RMS Error	R
Normal 5	.089	.745	.217	.285	.996
Normal 15	.091	.955	.095	.144	.996
Normal 25	.092	.971	.094	.140	.996
Rectangular 5	.093	.809	.170	.233	.996
Rectangular 15	.093	.955	.097	.146	.996
Rectangular 25	.086	.985	.089	.135	.996
Peaked 5	.043	.601	.324	.410	.996
Peaked 15	.062	.729	.225	.292	.996
Peaked 25	.081	.786	.184	.247	.996
No Linking	.003	.970	.125	.162	.996

root-mean-square errors in ability estimates was identical to that observed for the mean absolute error.

The correlations between true and asymptotic ability were uniformly .996 for the nine anchor tests, which is the same value observed when no linking was performed.

The summary characteristics of the asymptotic ability estimates for the heterogeneous case are presented in Table 57. These summary statistics had much the same pattern as those of the homogeneous case. As in the homogeneous case, the peaked tests produced estimates with means closer to zero than did the other anchor tests; these means increased with increased test length. The means for the normal and rectangular tests were essentially the same, and clustered between .083 and .090; they did not vary systematically with test size. The standard deviations of ability estimates were smallest for the peaked tests. They were closest to 1.0 for the rectangular tests, although the standard deviations for the normal tests were only slightly lower.

The no-linking condition produced estimates with a mean of -.013, closer to zero than any of the anchor tests. The standard

Table 57. Asymptotic Ability Estimates--Anchor Tests Heterogeneous Condition Using Systematically Sampled Examinees

Method	Mean	SD	Absolute Error	RMS Error	R
Normal 5	.086	.742	.216	.284	.996
Normal 15	.089	.951	.091	.136	.996
Normal 25	.089	.967	.091	.132	.996
Rectangular 5	.090	.806	.167	.231	.995
Rectangular 15	.090	.951	.092	.138	.996
Rectangular 25	.083	.982	.085	.126	.996
Peaked 5	.041	.598	.325	.411	.996
Peaked 15	.060	.726	.226	.292	.996
Peaked 25	.079	.782	.183	.245	.996
No Linking	-.013	.962	.095	.127	.995

deviation of estimates from the no-linking condition was .962; this was exceeded only by the 25-item normal and rectangular tests.

As before, the peaked tests performed most poorly in terms of mean absolute error, with values ranging from .183 to .325. The rectangular test performed slightly better than the normal test, although differences were small at the longer test lengths. At test lengths of 15 or larger, mean absolute error was less than .092 for both the normal and rectangular tests; these were the only tests with mean absolute error below the .095 observed for the no-linking condition. Mean-absolute error decreased with increased test length.

The pattern for root-mean-square error was similar. The peaked tests performed most poorly, with root-mean-square error from .245 to .411. The rectangular tests performed only slightly better than the normal tests, particularly at the longer test lengths. Under the no-linking condition, root-mean-square error was .127, which was exceeded by all tests except the 25-item rectangular test.

The correlation between true and asymptotic ability was .996 in all cases but one; when no linking was done, this correlation was .995.

Efficiency of ability estimation. The relative efficiencies of the various anchor test linking procedures for the homogeneous case are presented in Table 58. The average item information with the true item parameters was .314. This dropped to .278 with the estimated item parameters and, hypothetically, perfect linking.

Table 58. Efficiency Analysis--Anchor Tests
Homogeneous Condition Using Systematically Sampled Examinees

	Average	Efficiency Relative to	
Method	Item Information	True Parameters	Estimated Parameters
True Parameters	.314		
Est. Parameters	.278	.887	
Normal 5	.274	.875	.986
Normal 15	.275	.877	.988
Normal 25	.275	.877	.988
Rectangular 5	.274	.873	.984
Rectangular 15	.275	.876	.987
Rectangular 25	.275	.876	.987
Peaked 5	.274	.875	.986
Peaked 15	.275	.876	.987
Peaked 25	.275	.876	.987
No Linking	.266	.849	.957

The efficiencies of these linking methods, relative to that achieved by using true parameters, clustered between .873 and .887, with the highest figures observed for the normal tests. With respect to the estimated parameters, the efficiencies of these anchor tests ranged from .984 to .988, with the normal tests being slightly superior to the rest. All these values were higher than the .957 observed in the no-linking condition.

The relative efficiencies of the various anchor test linking procedures are presented in Table 59 for the heterogeneous case. The average item information with the true item parameters was .305. This dropped to .271 with the estimated item parameters and perfect linking.

Table 59. Efficiency Analysis--Anchor Tests
Heterogeneous Condition Using Systematically Sampled Examinees

Method	Average Item Information	Efficiency Relative to True Parameters	Efficiency Relative to Estimated Parameters
True Parameters	.305		
Est. Parameters	.271	.889	
Normal 5	.261	.858	.965
Normal 15	.262	.860	.968
Normal 25	.262	.859	.967
Rectangular 5	.261	.855	.962
Rectangular 15	.261	.858	.966
Rectangular 25	.262	.859	.966
Peaked 5	.261	.857	.964
Peaked 15	.262	.858	.966
Peaked 25	.262	.859	.967
No Linking	.248	.814	.916

The efficiencies of these linking methods, relative to that achieved by using true item parameters, clustered between .855 and .860. Once again, slightly higher figures were observed for the normal tests. With respect to the estimated parameters, the efficiencies of these nine anchor tests ranged from .962 to .968, with the normal tests being slightly superior to the rest. All these values were higher than the .916 observed in the no-linking condition.

Results--Robust-Maximum-Likelihood Scores

In addition to the Bayesian ability estimates which were computed for all simulated examinees, maximum-likelihood estimates were computed for the examinees included in the calibration groups of 1000. Identical analyses of item parameter error, asymptotic ability estimates, and efficiency were computed for these estimates for the homogeneous condition. For direct comparison with the results obtained using the Bayesian scores, summary statistics for the Bayesian scores were recomputed using only the 1,000-examinee calibration groups.

Fidelity of parameter estimation. Table 60 presents the combined results of item parameter error for the maximum-likelihood and Bayesian scores. For the maximum-likelihood scores, biases in the means of the a parameters were largest for the peaked tests and smallest for the rectangular tests although, again, differences between the normal and rectangular tests were small. All of the anchor tests except for the shortest two peaked tests, yielded smaller (in absolute value) bias figures than did the no-linking condition. Bias in the mean of the a parameters decreased with increased test lengths for the peaked tests, but no trends were observed with test lengths for the other anchor tests.

The bias in the standard deviation of the a parameters was of approximately the same magnitude for all three anchor test types, and showed no consistent trends with test lengths. The no-linking condition yielded a bias of .112, which was exceeded only by the 5-item tests.

With respect to the Bayesian scores, the largest bias in the mean of the a parameters was also observed for the peaked tests, the smallest bias for the rectangular tests. In general, bias figures were larger for the Bayesian scores. Biases for the standard deviations of the a parameters for the Bayesian scores, however, were of approximately the same magnitude as those observed for the maximum likelihood scores, although the maximum-likelihood scores yielded somewhat smaller bias for the peaked tests.

For the maximum-likelihood scores, the biases in the means of the b parameters were largest for the peaked tests, with small differences between the normal and rectangular tests. All of the bias values were larger than the .147 observed in the no-linking condition, although they all decreased with increased test lengths. Biases in the standard deviation of the b parameters were largest for the peaked tests, and again, there were only small differences between the normal and rectangular tests. These values decreased with increased test length, and all were greater than the .228 observed with no linking.

Table 60. Item Parameter Error--Anchor Tests
Homogeneous Condition Using Systematically Sampled Examinees

	Bayesian				Maximum Likelihood			
	Bias in		RMS		Bias in		RMS	
Method	Mean	SD	Error	R	Mean	SD	Error	R
Normal 5								
a	.575	.264	.906	.493	-.035	.248	.822	.329
b	.114	-.091	.338	.980	.453	.599	.962	.946
Normal 15								
a	.101	.100	.586	.489	-.003	.069	.594	.472
b	.217	.258	.506	.980	.232	.353	.535	.981
Normal 25								
a	.073	.089	.578	.489	.045	.081	.606	.479
b	.222	.281	.517	.980	.217	.300	.488	.982
Rect. 5								
a	.399	.202	.767	.491	.050	.191	.687	.423
b	.149	.018	.350	.980	.285	.439	.740	.955
Rect. 15								
a	.095	.096	.584	.492	-.022	.066	.606	.474
b	.219	.260	.497	.980	.249	.381	.560	.981
Rect. 25								
a	.043	.080	.566	.491	.037	.079	.598	.479
b	.227	.314	.544	.980	.213	.308	.490	.982
Peaked 5								
a	1.087	.447	1.384	.496	-1.047	-.185	1.182	.319
b	-.007	-.324	.419	.980	1.964	4.508	5.075	.954
Peaked 15								
a	.620	.281	.945	.494	-.688	-.075	.880	.370
b	.072	-.116	.32[illegible]	.980	1.100	2.050	2.508	.943
Peaked 25								
a	.457	.226	.811	.492	.017	.074	.599	.467
b	.123	-.017	.348	.980	.337	.327	.583	.980
No Linking								
a	.143	.112	.629	.501	.143	.112	.629	.501
b	.147	.228	.444	.973	.147	.228	.444	.973

Biases for the Bayesian scores were smaller, in general, than they were for the maximum-likelihood scores. They tended to increase with increased test lengths, and approximately half were smaller than the values observed with no-linking.

For the maximum-likelihood scores, root-mean-square error in the a parameters was largest for peaked tests. The advantage of the rectangular tests was slight. There was no consistent trend with test length; about half of the values were smaller than the value of .629 observed with no-linking.

This same pattern of root-mean-square errors in the a parameters was observed for the Bayesian scores, and the magnitude of the errors was approximately the same for the two scoring methods.

Root-mean-square errors in the b parameters for the maximum-likelihood scores were largest for the peaked tests, and the normal and rectangular tests performed equally well. There was a strong tendency for the root-mean-square error to decrease with increased test length, although all values were larger than the .444 observed with no-linking.

For the Bayesian scores, root-mean-square errors increased with test length for the normal and rectangular tests; the magnitude of the errors was much smaller for the Bayesian scores than for the maximum-likelihood scores.

The correlations between the true and estimated a parameters were smallest for the peaked tests and largest for the rectangular tests when using the maximum-likelihood scores. When the Bayesian scores were used, all the anchor tests produced correlations which were of approximately the same magnitude, and consistently higher than those observed for the maximum-likelihood scores.

For the maximum-likelihood scores, the correlations between true and estimated b parameters were of about the same magnitude for all the anchor tests, with the 15-item peaked test performing worse than would otherwise have been expected. For the Bayesian scores, these correlations were uniformly .980 for all nine anchor tests.

Characteristics of asymptotic ability estimates. Table 61 presents the summary statistics for the asymptotic ability estimates with maximum-likelihood and Bayesian scoring. When maximum-likelihood scores were used, the 5-item normal and all of the peaked anchor tests produced means somewhat deviant from zero. The remaining anchor tests produced means near .1. The no-linking procedure produced a mean of .034, better than that produced by any of the linking procedures.

The linking procedures did a better job of producing estimates with a mean of zero when these estimates were scores computed with a

Table 61. Asymptotic Ability Estimates--Anchor Tests
Heterogeneous Condition Using Systematically Sampled Examinees

	Bayesian				Maximum Likelihood			
Method	Mean	SD	RMS Error	R	Mean	SD	RMS Error	R
Normal 5	.092	.739	.290	.996	.225	1.027	.285	.995
Normal 15	.091	.945	.143	.996	.098	1.023	.147	.996
Normal 25	.092	.962	.138	.996	.098	.995	.147	.996
Rect. 5	.092	.805	.235	.996	.107	.965	.146	.997
Rect. 15	.093	.950	.143	.996	.117	1.044	.172	.996
Rect. 25	.084	.979	.130	.996	.088	.997	.138	.996
Peaked 5	.040	.597	.412	.996	.259	2.694	1.781	.980
Peaked 15	.058	.723	.295	.996	.490	1.796	.956	.997
Peaked 25	.079	.780	.249	.996	.204	1.008	.233	.996
No Linking	.034	.962	.133	.996	.034	.962	.133	.996

modal Bayesian algorithm. No mean was larger than .093. This was not surprising since the Bayesian algorithm explicitly regressed estimates toward zero. Again, there were but slight differences between the normal and rectangular tests. This time, however, the peaked tests performed best, with means between .040 and .079. Even these, however, were still larger than that obtained by not linking at all. Neither data set revealed a trend toward decreasing means with increased test length.

The normal and rectangular tests, coupled with maximum-likelihood scoring, produced estimates whose standard deviations were close to 1.0, typically between .965 and 1.044, with slightly "better" estimates produced using the normal tests. The peaked tests produced estimates with standard deviations quite large, at least for the 5- and 15-item tests. The longest peaked test, and all the normal and rectangular tests, produced estimates with standard deviations closer to 1.0 than was observed with no-linking.

With the Bayesian scores, ability estimates were systematically less variable, as would be expected from a procedure which regressed all estimates away from the extremes. The peaked test produced estimates less variable than the others; no standard deviation here was greater than .780. Although the differences were minor, the rectangular test produced estimates with standard deviations closer to 1.0 than did the normal test. Still, the no-linking value of .962 was exceeded only by the 25-item rectangular test.

There were few differences between the scoring procedures in terms of mean absolute and root-mean-square errors. For both procedures, the normal and rectangular tests performed best, with a slight advantage given to the rectangular test. Overall, the Bayesian scores performed slightly better than did the maximum-likelihood scores. In both cases, the peaked tests performed worst, although here the difference was much more marked for the maximum-likelihood scores. Only for the 25-item rectangular test with Bayesian scores did the errors ever drop below the level observed with no-linking.

All the correlations between true and estimated ability clustered near .996 when Bayesian scoring was used. These correlations were more variable with maximum-likelihood scoring and, for the peaked and rectangular anchor tests, showed a slight decrease with increasing anchor-test length.

Efficiency of ability estimation. Table 62 presents the efficiency figures for the maximum-likelihood and Bayesian scores. For the Bayesian estimates, average item information was essentially .267 for all nine anchor test conditions. For the maximum-likelihood scores, this level was not reached until the 15-item normal and rectangular anchor tests were used; for the peaked test, 25 items were necessary. For the Bayesian scoring, efficiencies were essentially the same for the three anchor test types, and these values increased only slightly with test length. All were above the level achieved in the no-linking condition. For the maximum-likelihood scores, the efficiencies were generally lower than for the Bayesian scores, even at the longest test lengths. All of the 5-item tests performed poorly, as did the 15-item peaked test. Efficiency, with respect to the estimated parameters, increased with test length, but still half the tabulated entries were below the value of .964 achieved with no linking.

Discussion

The data on anchor-test linking methods can be summarized rather briefly since there were several distinct trends with few exceptions. In terms of parameter bias, the peaked tests performed most poorly, often yielding large errors in parameter and ability estimation. There were few consistent differences noted between the normal and rectangular tests, especially for longer tests, although at the shorter test lengths, the rectangular test was usually superior. Differences among the test types tended to fade when the criterion was no longer bias but was the correlation between true and estimated

Table 62. Efficiency Analysis--Anchor Tests
Homogeneous Condition Using Systematically Sampled Examinees

	Bayesian			Maximum Likelihood		
		Efficiencies Relative to			Efficiencies Relative to	
Method	Avg. Item Info.	True Params.	Est. Params.	Avg. Item Info.	True Params.	Est. Params.
True Params.	.306			.306		
Est. Params.	.270	.882		.270	.882	
Normal 5	.267	.872	.989	.235	.770	.873
Normal 15	.267	.873	.990	.266	.870	.987
Normal 25	.267	.874	.992	.267	.872	.989
Rect. 5	.266	.871	.988	.254	.831	.943
Rect. 15	.267	.873	.990	.265	.867	.983
Rect. 25	.267	.873	.990	.266	.870	.986
Peaked 5	.267	.871	.988	.227	.741	.841
Peaked 15	.267	.873	.991	.249	.813	.922
Peaked 25	.267	.874	.991	.266	.869	.986
No Linking	.260	.850	.964	.260	.850	.964

parameters or true and estimated ability. Differences among the test types also disappeared when their relative efficiencies were taken as the criterion.

Anchor test length was a salient factor when one investigated the errors of _a_-parameter and ability estimation. Across test types, there were only small differences observed between the 15- and the 25-item tests; the 5-item tests were typically much worse than the others. The trend toward decreasing errors with increasing test lengths was expected, but was observed only for the _a_ parameters. For the _b_ parameters, this trend was reversed, with smaller errors observed with the shorter tests.

The test length effects disappeared when correlations and efficiencies rather than biases and errors were considered.

When comparisons were made between the Bayesian and the maximum-likelihood scores, the former were consistently better based on all the criteria used in this research.

Conclusions

Data presented in this section of the report provided the first opportunity to compare all four linking methods. In an effort to avoid confusion, only data relevant to the conclusions drawn are presented. Since the parameter-error statistics bear little direct relation to the utility of the linked items, they will not be discussed.

In terms of capacity to produce an asymptotic metric with the correct mean, the anchor-group method was generally superior. In nearly all configurations investigated, the anchor-group method produced a mean correct to the second decimal place. The Bayesian equivalent-tests method produced the most deviant mean. Asymptotic means for each of the methods were essentially equivalent in the homogeneous and heterogeneous conditions.

The most accurate asymptotic standard deviations were produced by the anchor-test method. With a 25-item rectangular anchor test, it produced an asymptotic standard deviation within .015 of the true value. In less favorable configurations, however, it produced standard deviations .4 unit in error. The equivalent-tests procedure produced results nearly as good as the best anchor-test configuration. The equivalent-groups and anchor-group procedures produced results somewhat less accurate.

Using root-mean-square error as a composite error-of-metric index, the anchor-group and anchor-test methods produced the least error and were approximately equivalent. The equivalent-tests method produced the most error.

Viewed in terms of linking efficiency, the anchor-test method produced the most efficient item pools. Its efficiencies ranged from .986 to .988 in the homogeneous condition and from .965 to .967 in the heterogeneous condition. Configured properly, the anchor group procedure resulted in equivalent efficiencies, but with smaller groups, the efficiency dropped somewhat. The equivalent-tests method produced efficiencies slightly lower than the least efficient of the two anchor procedures. The equivalent-groups method, whose assumptions were violated by these data, produced efficiencies slightly lower than those of the equivalent-tests procedure.

Although not considered in the previous discussion, the no-linking condition should not be forgotten. In terms of errors in the asymptotic distribution, it produced parameters as good as those produced by the best of the other methods. Its efficiencies were somewhat lower than those of the equivalent-groups procedure, however.

Use of the maximum-likelihood scoring procedure with the anchor-group or anchor-test procedures did not seem to be warranted by the data. In addition to producing less efficient item pools than did the Bayesian scoring procedure, this procedure appeared to bias the asymptotic metric more severely. Since it was investigated primarily as a means of reducing bias in the metric, these results suggest that it is not a useful scoring procedure for linking in the environment investigated here.

Neither of the anchor methods were evaluated in the randomly sampled data set because their performance in that set was assumed to be equivalent to their performance in the systematically sampled data set. The same assumption was reasonable for the equivalent-tests method but that method was, nevertheless, evaluated in both sets and thus provides a test of the assumption. In this data set the equivalent-tests method produced parameters with root-mean-square errors of .356 and .231 in the homogeneous and heterogeneous conditions, respectively, and efficiencies of .971 and .949. In the randomly selected data set, corresponding values were .209, .143, .962, and .944. The asymptotic error statistics appeared somewhat smaller in the randomly sampled condition but the efficiencies were comparable.

Efficiencies for the Bayesian equivalent-groups procedure were .988 and .973 for the homogeneous and heterogeneous conditions, respectively. These efficiencies compare very favorably with .988 and .968, the best efficiencies obtained by any method in the systematically sampled data set. This suggests that, if examinees are randomly sampled from the population of interest, the Bayesian equivalent-groups procedure can produce item pools as efficient as any of the more complicated methods.

VI. LINKING WHEN EXAMINEES ARE SELECTED

Investigations of linking discussed in previous chapters were limited to populations that could, more or less, occur in nature. No explicit selection had been done in defining the population and the distributions of abilities were essentially symmetric. The research discussed in this section of the report dealt with a selected population. The examinee samples used were those of the selected data set described in an earlier section. Briefly, the upper two-thirds of a sample were selected, on the basis of number-correct scores, to simulate selection that occurs in Air Force recruits. The procedure was very similar to that used by Ree (1978).

The selected data set contained only one row of the matrix of test lengths and sample sizes corresponding to a sample size of 1,000. This restriction of the data set was done primarily to save computer costs since adequate data regarding the joint effects of test length and sample size had been collected and discussed in earlier sections of this paper. Since the entire matrix was not available, only the homogeneous analyses were done.

Equivalence Methods

Procedure

The equivalence linking procedures used on the selected data set were similar in form to those used in previous sections; the same equations were used to perform the linking. Because of findings of previous sections, however, only the modal Bayesian scoring method was used for equivalent-groups linking. The remaining five linking methods were not used. The equivalent-tests and no-linking procedures were the same as before.

Results

Fidelity of parameter estimation. Table 63 presents fidelity-of-estimation statistics for the homogeneous condition using selected examinees. Columns one and two present means and standard deviations of the true a and b parameters for the items used with the selected data set. As was the case with items used in previous data sets, no notable departures from the population values were observed.

Biases in the parameter estimates are presented in columns three and four. The a-parameter means were essentially unbiased for the equivalent-tests and no-linking procedures. The a parameters were underestimated by .335 units when the Bayesian equivalent-groups procedure was used. The equivalent-tests procedure produced b parameters

Table 63. Item Parameter Error--Equivalence Methods
Homogeneous Condition Using Selected Examinees

Method	True Mean	True SD	Bias in Mean	Bias in SD	Absolute Error	RMS Error	R
Equiv. Groups							
a	1.601	.501	-.335	-.008	.476	.624	.466
b	.176	1.340	-.530	.843	.893	1.102	.974
Equivalent Tests							
a	1.601	.501	-.015	.112	.444	.589	.458
b	.176	1.340	.051	.390	.456	.622	.968
No Linking							
a	1.601	.501	-.015	.112	.491	.651	.465
b	.176	1.340	-.378	.400	.522	.657	.975

with nearly the correct mean. The other two procedures produced underestimates of the *b* parameters.

The Bayesian equivalent-groups procedure produced *a* parameters with nearly the correct standard deviation. Standard deviations of the *a* parameters were slightly greater than the correct values for the other two methods. All linking procedures produced *b*-parameter standard deviations that were larger than those of the true parameters. The equivalent-groups procedure produced the largest standard deviations.

Columns five and six present absolute and root-mean-square errors of parameter estimation. Errors in *a*-parameter estimates were approximately equal for all methods. The equivalent-tests method produced the least error and the no-linking procedure produced the most. Errors in the *b* parameters were about equal for the equivalent-tests and no-linking procedures. The equivalent-groups procedure produced *b*-parameter errors substantially greater than those produced by the other procedures.

Correlations between true and estimated parameters are presented in the last column of the table. The equivalent-groups and no-linking procedures were trivially different in terms of this correlation. The equivalent-tests procedure produced correlations somewhat lower than the other two procedures.

Characteristics of asymptotic ability estimates. Table 64 presents statistics descriptive of asymptotic ability estimates. These

Table 64. Asymptotic Ability Estimates--Equivalence Methods
Homogeneous Condition Using Selected Examinees

Method	Mean	SD	Absolute Error	RMS Error	R
Equiv. Groups	-.813	1.565	.823	1.000	.996
Equivalent Tests	-.156	1.250	.265	.369	.996
No Linking	-.566	1.265	.566	.642	.996

statistics should be interpreted relative to a standard normal population even though the items were calibrated on a population distinctly different. The first column presents asymptotic means resulting from application of the items to a standard normal population. All procedures resulted in net underestimates of abilities. The equivalent-tests procedure produced the mean closest to the true value of zero, and the equivalent-groups procedure produced the one most deviant.

Asymptotic standard deviations are presented in the second column. All three linking procedures produced estimates that were quite deviant from the mean. The equivalent-groups procedure produced the most deviant estimates, however, and the other two methods produced estimates about equally deviant.

Absolute and root-mean-square errors of the asymptotic estimates are presented in columns three and four. The equivalent-tests procedure produced the least error, according to both statistics, and the equivalent-groups procedure produced the most error.

Column five presents correlations between true and asymptotic ability estimates. All three procedures resulted in correlations of .996, indicating that the regressions were about equally linear.

Efficiency of ability estimation. Table 65 presents calibration and linking efficiencies for the selected data set. As was true of corresponding tables in previous sections, columns two and three are simply manipulations of the data in column one and column three is most informative relative to linking efficiency. As can be seen from column three, linking efficiencies of the equivalent-groups and no-linking procedures were equal. The linking efficiency of the equivalent-tests procedure was somewhat lower.

Linking efficiencies were quite high for all methods. These figures are not, however, directly comparable to those from previous

Table 65. Efficiency Analysis--Equivalence Methods
Homogeneous Condition Using Selected Examinees

Method	Average Item Information	Efficiency Relative to True Parameters	Efficiency Relative to Estimated Parameters
True Parameters	.325		
Est. Parameters	.268	.824	
Equiv. Groups	.265	.814	.988
Equivalent Tests	.262	.807	.979
No Linking	.265	.814	.988

data sets because these figures represent averages of only four cells rather than the 12 represented in previous tables.

Anchor Group Method

Procedure

The anchor-group linking procedure used for the selected data set was essentially the same as that used for the systematically sampled data set. The modal Bayesian scoring procedure was used throughout this section, as the maximum-likelihood procedure demonstrated no distinct advantages in previous analyses. Details of the linking procedure were presented in the previous section and will not be repeated here.

Results

Fidelity of parameter estimation. Table 66 presents parameter error for the anchor-group design in the selected data set. Bias in the estimates of the mean a parameter was positive for the normal group (indicating overestimates) and slightly negative for the uniform group (indicating underestimates). Bias tended to decrease with increasing anchor group size for both normal and uniform groups. Bias in the standard deviation of the a parameters showed the same trends as the means. Bias tended to decrease with increasing anchor group size and was smaller for the uniform group than for the normal group. The no-linking condition very slightly underestimated the

Table 66. Item Parameter Error--Anchor Groups
Homogeneous Condition Using Selected Examinees

Method	True Mean	True SD	Bias in Mean	Bias in SD	Absolute Error	RMS Error	R
Normal 10							
a	1.601	.501	.220	.213	.536	.703	.466
b	.176	1.340	.063	.182	.306	.429	.972
Normal 30							
a	1.601	.501	.181	.192	.517	.682	.464
b	.176	1.340	.044	.205	.309	.429	.973
Normal 50							
a	1.601	.501	.163	.187	.505	.672	.465
b	.176	1.340	.060	.221	.315	.434	.974
Normal 100							
a	1.601	.501	.144	.179	.503	.666	.467
b	.176	1.340	.043	.243	.321	.440	.974
Uniform 10							
a	1.601	.501	.129	.184	.492	.657	.456
b	.176	1.340	.030	.262	.348	.508	.972
Uniform 30							
a	1.601	.501	-.010	.125	.448	.601	.461
b	.176	1.340	.065	.395	.425	.577	.974
Uniform 50							
a	1.601	.501	-.005	.123	.460	.609	.464
b	.176	1.340	.057	.388	.417	.548	.974
Uniform 100							
a	1.601	.501	-.015	.119	.459	.610	.467
b	.176	1.340	.055	.401	.425	.561	.974
No Linking							
a	1.601	.501	-.015	.112	.491	.651	.465
b	.176	1.340	-.378	.400	.522	.657	.975

a-parameter mean and showed less bias in the *a*-parameter standard deviations than did any of the linking methods.

The biases in the means of the *b* parameters were very much alike for both anchor groups, but the no-linking condition substantially underestimated the mean. Bias in the standard deviation of the *b* parameters revealed a tendency for increasing bias with increasing anchor group size for both normal and uniform groups. The normal group, however, showed smaller bias in standard deviation than the uniform group, while the no-linking method had one of the largest biases in standard deviation.

Absolute and root-mean-square error for the [illegible] parameter showed a decreasing trend with increasing anchor group size for the normal groups. The uniform groups showed less error than the normal groups overall. The no-linking group showed errors midway between the uniform and normal groups.

Errors in the *b* parameters followed the opposite trends noted for the *a*-parameter errors; errors increased with increasing anchor group size and error was less for uniform groups than for normal groups. The no-linking group showed the greatest *b*-parameter error.

Correlations between true and estimated parameters tended to increase with increasing anchor group size and to be somewhat higher in the normal groups than in the uniform groups for the *a* parameter. For the *b* parameters, there were negligible differences between the groups. The correlation between true and estimated *a* parameters in the no-linking group was comparable to that observed in the normal and uniform groups and the *b*-parameter correlation in the no-linking group was the highest of all groups.

Characteristics of asymptotic ability estimates. Table 67 presents descriptive statistics for asymptotic ability estimates for each anchor group in the selected data set. Column one, showing the means, indicates that parameters linked using normal or using uniform anchor groups tended to underestimate the population mean of zero. The normal groups appeared to have closer estimates than the uniform groups over all group sizes, while the no-linking condition showed the greatest deviation from zero. There were no apparent trends with respect to increasing anchor group size.

Standard deviations were somewhat higher than the population value of 1.0 and showed a trend for increasing values as the anchor group size increased. The normal groups produced standard deviations closer to 1.0 than did the uniform groups, and the no-linking condition produced the largest standard deviation.

Absolute and root-mean-square error, presented in columns three and four, showed a tendency to increase with increasing anchor group size and to be larger for uniform than for normal groups. No-linking produced the largest errors.

There were no differences across group composition or group size in terms of the correlation of the true with the asymptotic ability

Table 67. Asymptotic Ability Estimates--Anchor Groups
Homogeneous Condition Using Selected Examinees

Method	Mean	SD	Absolute Error	RMS Error	R
Normal 10	-.084	1.081	.119	.161	.996
Normal 30	-.109	1.111	.130	.185	.996
Normal 50	-.094	1.118	.128	.181	.996
Normal 100	-.118	1.131	.143	.203	.996
Uniform 10	-.143	1.146	.168	.236	.996
Uniform 30	-.130	1.241	.217	.295	.996
Uniform 50	-.136	1.236	.217	.294	.996
Uniform 100	-.138	1.244	.222	.299	.996
No Linking	-.566	1.265	.566	.642	.996

estimates. All correlations, including the no-linking group, were uniformly .996.

Efficiency of ability estimation. Table 68 presents the average item information and relative efficiencies for the anchor-group linking method. The efficiencies relative to the estimated parameters, shown in column three, revealed a slight tendency to increase as anchor group size increased. The normal groups showed an almost trivial advantage over the uniform groups, while the no-linking condition showed the highest efficiency.

Discussion

Much of the information presented thus far has been less than definitive. Different analyses suggested different interpretations. Fidelity analyses, for example, suggested that anchor groups using a uniform distribution yield less parameter error than those using a normal distribution. Asymptotic ability statistics suggested that a normally distributed sample yields results superior to those of a uniform distribution. Efficiency analyses, on the other hand, showed both normal and uniform anchor groups to have about the same efficiency.

Table 68. Efficiency Analysis--Anchor Groups Homogeneous Condition Using Selected Examinees

Method	Average Item Information	Efficiency Relative to True Parameters	Efficiency Relative to Estimated Parameters
True Parameters	.325		
Est. Parameters	.268	.824	
Normal 10	.263	.810	.983
Normal 30	.265	.813	.987
Normal 50	.265	.813	.987
Normal 100	.265	.813	.987
Uniform 10	.263	.809	.982
Uniform 30	.263	.810	.983
Uniform 50	.263	.810	.983
Uniform 100	.264	.812	.986
No Linking	.265	.814	.988

Results of the efficiency analysis for the anchor-groups procedure were especially noteworthy in view of the rather large discrepancy between the distributions of ability used in the anchor groups and those used in the calibration samples. The anchor groups had abilities with a mean of zero and a standard deviation of one. The selected examinees in this data set had a mean greater than zero and a standard deviation less than one.

Although the no-linking condition showed the highest efficiency, the b-parameter mean and asymptotic ability mean were quite deviant from their true values. The reason the efficiency of the no-linking condition did not reflect these deviant parameter estimates is because efficiency statistics, like correlations, are insensitive to linear transformations of the data. If, however, an attempt was made to link items calibrated on groups widely different in ability (vertical equating), the no-linking procedure would show much lower efficiencies because each set of items would tend to shift the scale closer to its own metric.

As discussed earlier, efficiency analyses are the most appropriate evaluative criteria to apply to the linking procedures. The efficiency analyses suggested the following observations: (a) group composition tended to make very slight differences in observed efficiency, (b) there was a tendency for higher efficiency as test length increased and anchor group size increased, the latter being less pronounced than the former, and (c) increasing anchor group size did not substantially increase the efficiency.

Anchor Test Method

Procedure

The anchor-test linking procedures used for the selected data set presented in this section were identical to those used for the randomly and the systematically sampled data sets. Details of these linking procedures were presented earlier and will not be repeated here. Analyses were performed only for the condition where the items were originally calibrated on 1,000 cases for four different test lengths. Only the homogeneous condition is presented here. Modal Bayesian ability estimates were used throughout.

Results

Fidelity of parameter estimation. Fidelity-of-estimation statistics for the homogeneous condition are presented in Table 69. All of the anchor test procedures overestimated the $\underline{a}$ parameters, although this bias systematically decreased with increased anchor-test lengths. The smallest biases in the mean of the $\underline{a}$ parameters were observed for the rectangular tests, although at the longer test lengths the normal tests produced biases nearly as small. Much larger biases were observed for the peaked tests at all three test lengths. When no linking was performed on the data, bias in the mean of $\underline{a}$ parameters was -.015. This figure was exceeded by all nine anchor test methods.

Biases in the standard deviations of the $\underline{a}$ parameters were largest for the peaked tests. There were few differences observed in the biases for the normal and rectangular tests. All the biases systematically decreased with increased test length. In the no-linking condition, bias in the standard deviation of the $\underline{a}$ parameters was .112. This figure was exceeded by all nine anchor test methods.

All anchor test methods produced $\underline{b}$-parameter estimates that were essentially unbiased in their means. The largest bias observed, -.082, was quite small. The no-linking group produced considerable bias, by comparison. This was expected, however, as the mean ability levels of the calibration groups were substantially above zero.

Table 69. Item Parameter Error--Anchor Tests
Homogeneous Condition Using Selected Examinees

Method	True Mean	True SD	Bias in Mean	Bias in SD	Absolute Error	RMS Error	R
Normal 5							
a	1.601	.501	.617	.366	.794	.998	.466
b	.176	1.340	-.030	-.088	.262	.353	.973
Normal 15							
a	1.601	.501	.181	.194	.514	.672	.466
b	.176	1.340	.037	.219	.317	.450	.973
Normal 25							
a	1.601	.501	.156	.188	.506	.662	.467
b	.176	1.340	.050	.241	.329	.464	.973
Rectangular 5							
a	1.601	.501	.552	.337	.744	.939	.466
b	.176	1.340	-.007	-.054	.252	.344	.974
Rectangular 15							
a	1.601	.501	.188	.197	.518	.677	.466
b	.176	1.340	.044	.211	.313	.445	.973
Rectangular 25							
a	1.601	.501	.123	.174	.493	.646	.467
b	.176	1.340	.055	.273	.347	.489	.973
Peaked 5							
a	1.601	.501	1.192	.588	1.273	1.541	.465
b	.176	1.340	-.082	-.346	.344	.462	.973
Peaked 15							
a	1.601	.501	.748	.416	.896	1.113	.465
b	.176	1.340	-.033	-.157	.271	.367	.973
Peaked 25							
a	1.601	.501	.566	.345	.755	.951	.466
b	.176	1.340	-.002	-.057	.257	.353	.973
No Linking							
a	1.601	.501	-.015	.112	.491	.651	.465
b	.176	1.340	-.378	.400	.522	.657	.975

As was observed for the b-parameter means, all three peaked tests underestimated the b-parameter standard deviations; this bias decreased with increased test length. Biases in the standard deviation of the b parameters were of approximately equal magnitude for the normal and rectangular tests. Except at the 5-item test lengths, this bias was positive; for both the normal and rectangular tests, bias increased with test length. All of the anchor tests produced biases smaller than that observed for the no-linking condition.

Mean absolute and root-mean-square errors in the parameters are presented in columns five and six of Table 69. The peaked anchor tests performed most poorly according to both of these indices of error for the a parameters. In general, errors for the rectangular tests were smaller than for the normal tests although, as before, these differences were small. Both indices of error decreased with increased test length. In most cases, the no-linking condition yielded smaller absolute and root-mean-square errors in the a parameters than did any of the anchor test conditions.

Overall, the magnitude of absolute and root-mean-square errors in the b parameters was approximately equivalent for all three types of anchor tests. Both types of errors decreased with increased test length for the peaked tests, but increased with test length for the normal and rectangular tests. The no-linking procedure yielded larger absolute and root-mean-square errors in the b parameters than did any of the anchor-test methods.

The anchor-test-method correlations between true and estimated a parameters clustered between .465 and .467; for the no-linking condition, this value was .465. The anchor-test correlations for the b parameters were almost uniformly .973 (the correlation for the 5-item rectangular test was .974), slightly lower than the value of .975 observed with no linking.

Characteristics of asymptotic ability estimates. Table 70 presents the summary characteristics of asymptotic ability estimates for the homogeneous case. Columns one and two present the means and standard deviations of the asymptotic ability metric. All of the anchor tests produced means slightly below the targeted zero. None of the three test types produced means consistently closest to zero but the normal tests consistently produced means most deviant. Differences among these means were small, however. Means consistently decreased with test length for the rectangular tests and increased for the others. The no-linking procedure produced a mean much more deviant from zero than did any of the anchor-test methods.

All of the peaked tests produced ability estimates with standard deviations less than 1.0. The 5-item normal and rectangular tests did likewise. The longer normal and rectangular tests produced estimates with standard deviations greater than 1.0. In all cases, the

Table 70. Asymptotic Ability Estimates--Anchor Tests
Homogeneous Condition Using Selected Examinees

Method	Mean	SD	Absolute Error	RMS Error	R
Normal 5	-.117	.892	.135	.188	.996
Normal 15	-.115	1.111	.130	.195	.996
Normal 25	-.107	1.126	.133	.198	.996
Rectangular 5	-.102	.918	.115	.164	.996
Rectangular 15	-.107	1.105	.125	.186	.996
Rectangular 25	-.110	1.148	.146	.215	.996
Peaked 5	-.116	.709	.230	.325	.996
Peaked 15	-.106	.843	.145	.213	.996
Peaked 25	-.097	.913	.113	.165	.996
No Linking	-.566	1.265	.566	.642	.996

standard deviations of ability estimates increased with anchor test length. The standard deviation of the no-linking condition was 1.265, a value further from 1.0 than was produced by any of the anchor tests.

Mean absolute and root-mean-square errors in the ability metric are presented in columns three and four of Table 70. The magnitude of absolute error was approximately the same across the three types of anchor tests, with a tendency for the smallest peaked test to produce errors larger than the rest. Mean absolute errors increased with test length for the rectangular tests, and decreased with test length for the peaked tests. For the normal tests, these errors did not vary systematically with test length. Mean absolute error in the no-linking condition was much higher than that observed for any of the anchor tests. Exactly the same patterns were observed for the root-mean-square errors in the ability estimates.

The correlation between true and estimated ability was uniformly .996 for all the anchor tests and for the no-linking procedure.

Efficiency of ability estimation. Information and the relative efficiencies for the anchor-test procedures for the homogeneous case are presented in Table 71. The average item information with the true parameters was .325. This dropped to .268 with the estimated parameters and, hypothetically, perfect linking. The average item information with the anchor-test procedures and with no-linking was .265.

Table 71. Efficiency Analysis--Anchor Tests
Homogeneous Condition Using Selected Examinees

Method	Average Item Information	Efficiency Relative to True Parameters	Efficiency Relative to Estimated Parameters
True Parameters	.325		
Est. Parameters	.268	.824	
Normal 5	.264	.813	.987
Normal 15	.265	.815	.989
Normal 25	.265	.814	.988
Rectangular 5	.265	.815	.989
Rectangular 15	.265	.815	.989
Rectangular 25	.265	.814	.988
Peaked 5	.265	.814	.988
Peaked 15	.265	.814	.988
Peaked 25	.265	.814	.988
No Linking	.265	.814	.988

The efficiencies of these linking methods, relative to that achieved by using true parameters, clustered between .813 and .815. With no linking, the relative efficiency was .814. With respect to the estimated parameters, the efficiencies of the anchor test procedures ranged from .987 to .989, with no overall difference observed across anchor tests. The corresponding efficiency figure for the no-linking condition was .988.

Discussion

Overall, the peaked anchor tests tended to perform most poorly when errors in item parameters were taken as the criteria. There were few differences observed between the normal and rectangular tests but, when differences were found, they tended to favor the rectangular tests. In most cases, the indices of bias decreased with increased test length; the 15-item tests performed nearly as well as the 25-item tests and better than the 5-item tests. There were essentially no differences across anchor test types and test lengths in the correlations between true and estimated item parameters.

More relevant to the study of linking methods are the characteristics of the asymptotic ability estimates produced by each method. There were few differences observed across anchor test types in terms of their ability to produce estimates with a mean of zero and standard deviation of one, and in the absolute and root-mean-square errors in these estimates. When differences were found, they typically indicated that the peaked tests were somewhat worse than the others. There were no consistent trends with test length. The correlations between the true and estimated ability were identical across all nine anchor tests.

Perhaps most important in this study, however, were the indices of efficiency of the anchor test procedures. Essentially no differences were found across anchor test types and test lengths; all efficiency figures were between .987 and .989.

Conclusions

Analyses presented in this section have been, in part, a replication of analyses done on the randomly sampled examinees. Examinees used in this section were randomly sampled from a single population. The difference between these groups and those of the previous data set was simply that the single population was redefined as having been selected, and thus skewed in distribution.

Many of the findings with the selected sample paralleled those of the randomly sampled data set. Specifically, equivalent-groups or no-linking methods produced pools of items as efficient, in terms of linking, as did the more complex anchoring methods. The equivalent-tests method, as before, was inferior to the other methods.

The anchoring methods were far superior to the equivalence and no-linking methods in reproducing the original standard ability metric. This was simply due to the fact that only the anchoring methods had information regarding the "correct" metric.

As a general conclusion, it appears that the equivalent-groups method is simple and effective for linking sets of items if examinees used in calibration are all sampled from a common population, regardless of its shape. If, however, the original metric must be reproduced, the equivalent-groups method has no way to reproduce it. Mixing items calibrated on a selected group with items calibrated on an unselected group would be one example where an original, or at least a common, metric would need to be reproduced.

VII. PRACTICAL APPLICATIONS OF LINKING

Development of a Composite Approach

The linking tasks the Armed Services must face in developing adaptive-testing item pools can be reduced to two. First, the items comprising the initial pool will be calibrated in several sets on several groups and must be linked onto a common metric. Second, new items will be added to the pool at later dates and must be linked onto the same metric. Data presented in the preceding sections provide good solutions to the first problem. These solutions will be summarized below. Data presented in these sections provide some solutions to the second problem. More complex solutions, however, require further analyses. (See Appendix C for a summary of a meeting with Air Force personnel in which the Armed Services linking problem was discussed.)

The primary objective of linking is to produce a pool of items that will function together efficiently. Efficiency of the method is thus the most important criterion for choosing a method to link the initial pool. Since norms will undoubtedly be constructed on the basis of the metric of the initial pool, additional criteria must be considered in choosing a method for linking future items to the original pool. Specifically, addition of the new items should not distort the original metric and, therefore, a method that produces little distortion should be chosen. Hence, the asymptotic-estimate criteria are also relevant to this linking problem. Discussion and analyses presented below will be limited to these relevant criteria.

Linking the Initial Item Set--A Summary of Findings

Given that the objective in calibrating and linking the initial item pool is to obtain a set of items that function efficiently, several methodological suggestions can be made. The equivalent-groups linking method using modal Bayesian scoring works as well as any of the more complicated linking procedures when examinees are randomly sampled from a common population. If it is possible to sample in this manner, there is no advantage to using a more complicated procedure. The method worked about equally well at all test lengths investigated. It exhibited a slight tendency toward greater efficiency with larger examinee samples, but these findings were inconsistent. The differences were not sufficiently consistent to suggest whether 500, 1,000, or 2,000 examinees should be used; in practice, the largest available sample would probably be used.

Analyses of calibration efficiency provided some guidance regarding the sample size and test length necessary for item calibration. Generally, larger samples and longer tests produced more efficient parameter estimates. If a tradeoff could be made between test length and sample size, however, these analyses suggested that emphasis should be placed on increasing the test length, since increases in test length were three to four times as effective as proportionate increases in sample size.

In the Armed Services environment, it is conceivable that new test items might be calibrated in conjunction with AFEES administration of the current ASVAB. If the new items were to parallel a subtest on the ASVAB, this subtest would be a potential anchor test, but random distribution of experimental subtests across the AFEES population would eliminate the need for an anchor test. Simultaneous calibration of the new and old ASVAB items would, however, result in a longer test and, therefore, better calibration so the two tests should be calibrated together, even if the ASVAB subtest is not used for linking.

If random distribution were to prove impractical, the analyses of previous sections suggest that an anchoring method should be used. Either 100 anchor examinees or 15 to 25 anchor items would provide efficiency equivalent to that obtained by randomly sampling examinees. If the new items were to be administered concurrently with the ASVAB, the anchor-test method of linking would be an obvious choice. Previous analyses suggest that rectangular and normal anchor tests work about equally well. Each of the present ASVAB subtests has an information curve which is similar to one of these two forms.

Linking Across Time--Further Analyses

An item pool, regardless of the care taken in its creation, is not likely to remain static forever. For a variety of reasons, new items will be added and old items will be removed during the life of the item pool. These new items must be calibrated and linked onto the metric of the original items.

Since the examinee population is likely to change over time, the equivalent-groups procedure is not an appropriate method of linking the new items to the old. The equivalent-tests procedure, even if its assumptions could be met, would still be an inefficient procedure. Given that individuals are likely to change over time, the anchor-group procedure would not be appropriate.

The anchor-test method, if the anchor test remained constant, would be as efficient over time as it is at a single time. Therefore, it appears to be the method of choice for linking over time. If a constant anchor test can be maintained, linking over time will produce no more difficulty than linking within a single time period.

It is conceivable, however, to perform anchor-test linking using several anchor tests over time. A current ASVAB subtest may be used as an anchor test for new items. These new items may be used to form a new ASVAB subtest. This new ASVAB subtest may then be used as an anchor test for linking the second new set of items. Before this cascading procedure is attempted, however, it is important that its effects on efficiency and the ability metric be known. (This is probably an oversimplification of the problem since future versions of the ASVAB are likely to be adaptive. It provides a manageable model for analysis, however, and should provide some insight into the problem.)

Method. Item parameters and ability levels for a sample size of 1000 and test lengths of 20, 35, 50, and 65 items were taken from the systematically sampled data set. This data set was chosen because each group within each of the four cells was sampled from a different population. This is analogous, to some extent, to what would happen if groups were sampled at different time periods.

Within each cell, five calibration groups were arbitrarily ordered. The first group was linked, using the equivalent-groups procedure, to a standard (i.e., mean zero, variance one) population. (Note that this does not imply anchoring, and each initial group was linked to a different standard population.) Fifteen items were then selected from the test given to the first group as an anchor test. The first 15 were selected and, since the items in the tests were randomly ordered, represented a randomly sampled subset of items. These items were administered to the second calibration group and, using these items as an anchor test, the items in the second test were linked to the first. Fifteen items were selected from this linked second test and used to link the third test. This procedure was repeated until the fifth test had been so linked.

Asymptotic-ability-estimate and efficiency statistics were then calculated. They were calculated on the first test alone and then on each of the remaining tests in combination with the first. Cumulative effects of linking could thus be observed as more new tests were cascaded upon the old.

Although the modal Bayesian scoring procedure had proved superior to the maximum-likelihood procedure when a single anchor test was used, it was not obvious to what extent its inherent bias would affect linking in a cascaded environment. The robust-maximum-likelihood procedure was thus additionally considered as an unbiased procedure.

Results. Table 72 presents asymptotic-ability-estimate means and standard deviations for cascaded linking using modal Bayesian scoring. The level of linkage refers to the number of linkages required to link back to the original test. Average errors represent the average absolute deviation of the row or column entries from the

Table 72. Asymptotic Ability Metric of Cascaded Tests--Modal Bayesian Scoring

	Level of Linkage	Test Length				Average Error
		20	35	50	65	
Mean	0	.118	.488	.053	.154	
	1	.052	.434	.064	-.032	.079
	2	-.152	.337	.047	-.048	.157
	3	-.028	.279	-.027	-.034	.156
	4	.116	.329	-.009	.073	.076
	Average Error	.121	.143	.040	.164	.117
Standard Deviation	0	1.136	1.189	1.089	1.194	
	1	1.057	1.080	.936	.914	.155
	2	.893	.909	.912	.872	.256
	3	.854	.801	.943	.842	.292
	4	.949	.880	.918	.887	.244
	Average Error	.198	.272	.161	.315	.237

zero-level values. The zero-level values differ from each other because no anchor method was used to anchor the first tests to any common metric.

The most notable observation that can be made from the first half of Table 72 is that there were no apparent trends in error with increasing linkage distance at any of the four test lengths with respect to the means. The column with the most deviant starting value, .488, showed some tendency to drift toward zero but this trend was not consistent.

The standard deviations exhibited a tendency to drop with the first one or two linkages. After that they appeared to stabilize at approximately .9. No differences in this tendency were apparent across the various test lengths.

Table 73 presents asymptotic-estimate means and standard deviations for robust-maximum-likelihood scoring. Unlike the Bayesian procedure, the maximum-likelihood procedure showed a slight tendency to produce increasing means with increasingly distant linkages. This tendency was inconsistent, however.

Table 73. Asymptotic Ability Metric of Cascaded Tests--Maximum-Likelihood Scoring

	Level of Linkage	Test Length				Average Error
		20	35	50	65	
Mean	0	.079	.406	.048	.103	
	1	.061	.497	.070	.145	.043
	2	.120	.537	.062	.163	.062
	3	.225	.592	.040	.210	.112
	4	.174	.558	.044	.247	.100
	Average Error	.075	.140	.012	.088	.079
Standard Deviation	0	.876	.951	.906	1.018	
	1	.845	1.015	.945	1.121	.059
	2	1.009	1.026	.998	1.123	.101
	3	1.083	1.107	1.047	1.183	.167
	4	.995	1.073	1.038	1.232	.147
	Average Error	.123	.104	.101	.146	.119

Standard deviations, using the robust-maximum-likelihood procedure, rose rather than fell. By the third linkage, they were deviant from the initial values by .167, on the average. This dropped to .147 by the fourth linkage and may be indicative of a stabilization.

Table 74 presents linkage efficiencies of the cascaded tests using modal Bayesian scoring. No consistent trends in efficiency were observed. A slight inconsistent trend toward lower efficiency with increasing linkage distance and an inconsistent increasing trend with respect to test length were observed. The overall level of efficiency was somewhat lower than levels observed previously in the systematically sampled data set; efficiencies with Bayesian anchor-test linking using a constant anchor test were .970, compared to .929 here. It should be noted, however, that the conditions of linking were somewhat different as five tests at a time were linked before, and only two at a time were linked here.

Table 75 presents linkage efficiencies of the cascaded tests using robust-maximum-likelihood scoring. A more definite decreasing trend in efficiency with linkage distance was observed here than had been observed using Bayesian scoring. An inconsistent increasing

Table 74. Linkage Efficiency of Cascaded Tests--Modal Bayesian Scoring

Level of Linkage	Test Length 20	35	50	65	Average
1	.943	.981	.983	.930	.959
2	.874	.914	.954	.918	.915
3	.895	.862	.969	.911	.909
4	.958	.883	.959	.936	.934
Average	.918	.910	.966	.924	.929

Table 75. Linkage Efficiency of Cascaded Tests--Maximum Likelihood Scoring

Level of Linkage	Test Length 20	35	50	65	Average
1	.968	.962	.993	.972	.974
2	.972	.923	.989	.965	.962
3	.865	.892	.967	.940	.917
4	.920	.911	.972	.863	.917
Average	.931	.922	.980	.935	.942

trend with respect to test length was again observed. In general, the maximum-likelihood scoring procedure produced somewhat more efficient linkage than did the Bayesian procedure. Where the average linking efficiency was .929 for the Bayesian procedure, it was .942 when maximum-likelihood scoring was used.

Discussion. Linking using cascaded anchor tests with Bayesian scoring did not exhibit any substantial tendencies toward decreasing efficiencies with increasing linkage distances. Slightly more consistent tendencies toward lowered efficiency were observed with maximum-likelihood scoring. Maximum-likelihood scoring produced slightly higher average efficiency than did Bayesian scoring across the conditions investigated. Slight trends in bias were observed with respect to asymptotic standard deviations using either method but none were observed with respect to means or efficiencies.

It should be noted that no trends were built into the true abilities used in this simulation. Abilities of each group were different but not in any predictable fashion. If trends were present in the true abilities, a trend might be noted in the estimation errors. A substantial long-term trend in ability is unlikely to be observed in Armed Services testing, however. Short-term trends produced by a military draft situation are unlikely to affect more than one or two generations of test items. Such a situation is similar to the one simulated here.

Design for a Specific Application

Following is an example of how the information learned about linking techniques in the preceding sections could be applied to a practical linking problem such as might be faced by the Armed Services. The problem presented below is one developed, in cooperation with Air Force personnel, to be representative of the linking problem the Armed Services will encounter in the development of an item pool for computerized adaptive administration of the ASVAB or its successor. The problem described is presented only as a hypothetical linking environment. The test described, while intended to reflect expected conditions, is not based on specific studies and should not be considered optimal, in any sense, for test design.

Description of the Problem

A new adaptive version of the ASVAB is to be developed. It will contain 10 subtests, 8 of which will be power subtests. Only the power subtests will require calibration by IRT methods. For each of these eight subtests, a pool of approximately 200 items will be developed. These items will be similar to items previously used in the ASVAB, with the exception that they will be written to cover the difficulty range from $b = -2.5$ to $b = 2.5$. The distribution of difficulty is expected to be nearly rectangular with somewhat heavier representation in the center.

Examinees for use in calibration will come primarily from all the AFEES. One additional hour of examining time to take experimental tests will be provided for 1,000 examinees at each of the AFEES. This means that roughly 50 new items, on the average, can be administered along with the current ASVAB. The eight item pools, in total, will contain 1,600 items. If 65,000 examinees each take 50 items and the 1,600 items are equally apportioned, each item will be administered to 2,031 examinees.

Some of the new subtests will parallel subtests on the current ASVAB; others will not. It is not essential that all individuals within a given AFEES take the same test. It is essential that the administration instructions and time requirements be identical for all experimental tests given within a single AFEES.

The objective of calibration and linking of these items is to obtain eight item pools, each of which contains items which function efficiently together for estimating ability. The actual scale on which the items are linked is not critical but, if the new items parallel an old ASVAB subtest, there should be a way of translating the new test scores to the familiar ASVAB scores. Furthermore, there should be some provision by which new items can be added to a pool and linked to the original metric.

A Proposed Linking Design

When applicable, the equivalent-groups method of linking provides the most trouble-free and efficient linking available. It appears that tests can be randomly distributed among AFEES if care is taken and thus the equivalent-groups procedure is the method of choice. The Bayesian scoring procedure is the preferred scoring method.

Three major factors should be kept in mind when assembling the experimental tests. First, administrative constraints require that all tests use the same administration instructions and that each requires no more than an hour to complete. Second, calibration efficiency is enhanced with longer tests. Third, calibration of each pool in equal-sized sets of items on equal numbers of examinees results in greatest linking efficiency.

Prior to assembling the administration packets, rough time estimates for completion of items in each of the pools should be obtained either from pilot administration or from past experience. Each pool should then be divided into the largest equal parts that can be administered within the time allowed. No item overlap is required.

Examinees can be apportioned across the eight pools equally or unequally. If they are to be apportioned equally, the number of examinees can be decided by simply dividing 65,000 by the number of item subsets. It may be more appropriate, however, to apportion unequally. The number of examinees apportioned to each subtest may be decided by the relative importance of the pools, the relative ease of calibration of the various item types, the number of subtests within each of the item pools, or by other considerations. Samples used within a pool should be of equal size; samples for different pools do not need to be of equal size.

Experimental tests should be randomly distributed among AFEES (and their mobile testing sites). While data presented in preceding sections have suggested that the equivalent-groups procedure works reasonably well even when tests are systematically distributed, non-randomness may result in the equivalent-groups method being less efficient than one of the anchoring methods. If the items in a pool parallel an ASVAB subtest which is routinely administered to all examinees, the ASVAB items should be combined with each of the individual experimental tests when calibration is done. If distribution of test packets is done randomly, no explicit attempt at anchoring need be done; the purpose of including the ASVAB items is simply to increase calibration efficiency by increasing the test length. If distribution is non-random, explicit anchoring may be desirable.

Conceptually, expressing scores of the new tests in terms of the old ASVAB scores may seem to be a simple matter of using the appropriate ASVAB subtest as an anchor test and then anchoring new items to it. Ability estimates from the new tests should, it seems, be equivalent to ability estimates from the old. There are two reasons why this is not the case. For finite-length tests, regardless of the scoring procedure used, ability estimates will contain some error and be biased to at least a small degree. Unless the ability estimates from the ASVAB subtest and the new items have equivalent error and bias, ability estimates of one will not be equivalent to the other, even if linking is perfect. Secondly, the old ASVAB is not expressed in an IRT ability metric. Obviously, then, ability estimates from the old ASVAB will not be equivalent to ability estimates from the new tests, even for infinitely long tests.

So even after the item pools are linked, correspondence between the new adaptive ASVAB and the old conventional ASVAB will not be immediately available. These correspondences can be developed by conventional equating procedures but only after the item pools are incorporated into a testing strategy and its error properties are known.

Addition of new items to the pool at a later time will require an anchor test. The most straightforward choice for such a test is a conventional test composed of items from the original ASVAB or the original new item set and kept constant in composition for all future additions. Research in a previous section suggested, however, that new anchor tests can be selected as time passes with slight efficiency loss and little bias. Use of the new ASVAB as an adaptive anchor test is another possibility. Further research into adaptive anchor tests should be done before such a method is applied, however.

VIII. SUMMARY AND CONCLUSIONS

Summary

Previous Literature

This report began with a review of the psychometric literature relevant to linking and equating which resulted in a number of findings. The first was a general framework for classification of linking and equating designs. Linking methods were classified on two general aspects: the design by which data are collected and the algorithm by which the linking transformations are made. The data collection designs were of four types: (a) sampling of equivalent examinees (equivalent-groups method), (b) sampling of equivalent items (equivalent-tests method), (c) anchoring through a common group of examinees (anchor-group method), and (d) anchoring through a common set of items (anchor-test method). There were a variety of transformation algorithms which can be grouped into linear, nonlinear, and Item-Response-Theory (IRT) methods.

Since the overall research project was limited to linking of IRT-calibrated items, the review concentrated on IRT linking and equating studies. The vast majority of the reported studies used the Rasch IRT model. These tended to be more descriptive than evaluative. The more evaluative studies suggested that Rasch equating worked well for examinees of average or above average ability but worked poorly when low-ability groups were equated to higher-ability groups. This deficiency was probably due to the model's inability to handle guessing.

Among the studies investigating linking using the more appropriate three-parameter IRT models, there was some confusion regarding the distinction between prediction, linking, and equating. A distinction was made here by defining prediction as relating scores on one psychological dimension to scores on another using regression techniques, by defining equating as establishing a correspondence between two tests measuring the same psychological dimension using non-regression techniques, and by defining linking as putting parameters of items measuring the same psychological dimension on the same scale. Examples of research which inappropriately confounded these techniques were discussed.

Linking Criteria

The criteria used in past studies for evaluating the adequacy of calibration, linking, and equating were not only confusing but, typically, not useful for comparing various techniques. Two new classes of criteria were developed for use in this project. The first considered the asymptotic characteristics of ability estimates using estimated item parameters. Through this class of criteria, the biasing effects of calibration and linking errors could be assessed. The second class of criteria consisted of the information and relative efficiency of ability estimation resulting from the use of item parameters containing calibration and linking errors. These criteria were used to assess the relative test lengths required by the various methods to produce equivalent precision of measurement. Techniques for separating amounts of inefficiency due to calibration and to linking were presented.

Simulation Design

Considering deficiencies in previous studies of linking, a simulation study to determine appropriate linking methods was designed. In developing the simulation model, care was taken to ensure that the test items specified were similar (in terms of their item parameters) to Armed Services items likely to be encountered in actual linking problems, and that the populations of simulated examinees were defined to be similar in ability to those likely to take such tests.

Item parameters were specified after analysis of available data on current ASVAB forms. Included in these data were IRT item parameters for an experimental ASVAB form paralleling Form 7 and conventional item parameters from norming administrations of new ASVAB Forms 8, 9, and 10. The ability distributions were determined from samples of 500 examinees from each of 65 AFEES responding to ASVAB Form 7.

The distributions of both ability levels and item parameters were generated from the mean, variance, skew, and kurtosis of the AFEES or ASVAB distributions using a random number generator capable of generating distributions of shapes specified by these four moments. Three basic data sets were created. The first, the randomly sampled data set, contained five replications at each of 12 combinations of test length and calibration sample size and simulated the condition in which test booklets were randomly distributed among the entire AFEES population. The second, the systematically sampled data set, contained the same combinations of test length and sample size but simulated the condition in which test booklets were distributed systematically among relatively few AFEES. The third, the selected data set, contained only one sample size and simulated the condition in which a selected recruit population was used.

Three categories of evaluative criteria were used to assess the adequacy of calibration and linking. The first category, fidelity of estimation, examined the relationships between true and estimated item parameters. Statistical indices used included the familiar bias, absolute error, root-mean-square error, and correlation used in previous studies. The second category, characteristics of asymptotic ability estimates, examined the relationships between true and asymptotic (i.e., infinite-test-length) ability estimates. Statistical indices included the mean, standard deviation, absolute and root-mean-square error of the estimates, and the correlation between true and asymptotic ability. The last category, efficiency of ability estimation, included average item information (an index closely related to the precision of estimation) and relative efficiency, the ratio of information from two sources. In this study, efficiencies were computed relative to the true and estimated item parameters, yielding efficiency indices of the linked items and linking procedure, respectively.

Results

In evaluating the basic data sets, three general conclusions were reached. First, the parameter correlation data generally supported other studies which assessed the calibration effectiveness of OGIVIA, the calibration program used in this study. The $\underline{b}$ parameters were very well estimated and the $\underline{a}$ and $\underline{c}$ parameters were less well estimated. Second, test length appeared to be relatively more important to calibration effectiveness than was sample size; efficiency analyses suggested that increases in test length were at least three to four times as effective in improving calibration efficiency as proportionate increases in calibration sample sizes. Finally, there was little difference in calibration efficiency between randomly and systematically sampled examinees, but there was a large difference in efficiency between these and the selected examinee groups.

In the randomly sampled data set, two general linking methods, the equivalent-groups and the equivalent-tests methods, were evaluated and compared. Comparisons were done in both a homogeneous linking condition, where the items to be linked were calibrated in tests of equal length using examinee samples of equal size, and in a heterogeneous condition of mixed test lengths and examinee sample sizes.

The fidelity-of-parameter-estimation analyses were unable to provide any conclusive evidence regarding which linking procedure was most effective. The asymptotic ability analyses, however, suggested that two linking procedures based on Bayesian ability estimation (an equivalent-groups procedure) were somewhat more effective than the others and that the equivalent-tests method was typically no better than not linking at all. The third set of analyses, those using the relative efficiency criteria, suggested that the equivalent-groups procedures were superior to the equivalent-tests procedures and that those using Bayesian scoring procedures were slightly superior to the others tested. Relatively little efficiency was lost when the OGIVIA-produced parameters were used with no explicit linking. Efficiency loss due to linking error was always less than that due to calibration error and, although test length and sample size had a definite effect on calibration efficiency, no strong effects appeared with respect to linking efficiency.

In the systematically sampled data set, two additional linking methods were considered along with the equivalence methods. The anchor-group method linked item sets using common examinee groups of different sizes and compositions. The anchor-test method linked item sets using common tests of different sizes and compositions. In terms of linking efficiency, the anchor-test method produced the most efficient item pools. The anchor-group method resulted in efficiencies equivalent to those of the anchor-test procedure if large groups were used, but with smaller groups the efficiencies dropped somewhat. The equivalence methods were somewhat less efficient than either of the anchor methods. Bayesian scoring was the method of choice. Maximum likelihood appeared not to be a useful scoring procedure for the linking conditions investigated.

Results from analyses based on data from linking when examinees were selected tended to parallel those of the randomly sampled data set. The equivalent-groups and no-linking methods produced item pools as efficient as the more complex anchoring methods. These methods were ineffective in recovering the original metric, however. Mean asymptotic estimates were biased downward considerably from the true values, and standard deviations were larger than the true values. One of the more complex methods would have to be used if recovery of the original metric was desired.

Application to a Practical Linking Problem

An application of the results of this research to a practical linking problem was described. The problem consisted of calibration and linking of item pools for computerized adaptive administration of the ASVAB. The general suggestion was that experimental test booklets be randomly distributed and equivalent-groups linking be used. For addition of items at later times, an anchor-test linking method was suggested. A further simulation was done to investigate the effect of cascaded anchor tests in which a new anchor test was created for each link. Neither excessive drift nor loss in efficiency was noted. It was concluded that such cascading could be used if necessary but that a constant anchor test should be preferred. When maximum-likelihood and Bayesian scoring procedures were compared, in the cascaded condition, the maximum-likelihood procedure showed a slight efficiency advantage over the Bayesian procedure.

Conclusions

If the item-linking procedures suggested in this report are followed, parameter errors due to imperfect linking should be a relatively minor problem in the development of an adaptive-testing item pool. With proper procedures the efficiency loss due to linking errors should be approximately 1%. This is small in comparison to the 10% to 12% efficiency loss due to calibration errors. This study thus appears to have answered the question: How should different item sets calibrated in different examinee groups be linked?

Next to the findings regarding linking, perhaps the most important results of this project were the developments of new classes of criteria of calibration and linking adequacy. It is conceivable that calibration, noted to be a greater problem than linking, might be improved by using a different calibration program. Prior to this study, no adequate method of comparing calibration effectiveness of various calibration programs and algorithms had been available. The efficiency statistics presented here allow a direct comparison of various procedures in terms of their capacity to provide parameters conducive to accurate estimation of ability. Since ability estimation is the objective of ability testing, these criteria seem ideal.

Analyses of the basic data sets using the program OGIVIA were presented in sufficient detail that they could easily be replicated using other calibration techniques. Evaluation of other calibration techniques using the efficiency criteria should quickly answer the question of which procedure is best. Since efficiency has a direct translation into test length, it should be useful in a cost-benefit analysis of the various procedures if the best procedure also should turn out to be the most expensive.

The asymptotic-estimate criteria should have application in evaluating various equating methods. In this study, these criteria showed that, using estimates of the item parameters, the relationship between true and asymptotic ability was not perfectly linear. In populations such as those considered here, this did not appear to be a great problem. This nonlinearity may be a problem in the vertical equating of tests of widely different difficulty levels. It was not uncommon for tests investigated in this project to fail to yield ability estimates much below -2.0. If two tests were substantially different in difficulty and the parameters were less-than-perfect estimates, the relationship between the two tests might be nonlinear. This is an area that should be investigated before IRT vertically equated tests are used for real decisions.

As a third area for application of the new criteria, efficiency analysis might be applied to investigating the appropriate number of parameters in an IRT model. Rasch enthusiasts, and some others, have suggested that the Rasch model is the appropriate method to use because other parameters in the multi-parameter models are too difficult to estimate. Using efficiency analysis, it should be possible to determine how many examinees and items are required for the additional parameters in a two- or three-parameter model to produce a net gain in measurement efficiency.

In summary, it is likely that there will be few questions concerning the development of Armed Services adaptive testing pools that cannot be answered from data presented in this report. Calibration presents somewhat more of a problem than does linking, but further research using criteria developed here should help solve this problem. Finally, developments resulting from this project may aid in the solution of some other IRT-related psychometric problems.

REFERENCES

Andersen, E. B. A strictly conditional approach in estimation theory. *Skandinavisk Aktuarietidskrift*, 1971, *54*, 39-49.

Andersen, E. B. Sufficient statistics and latent trait models. *Psychometrika*, 1977, *42*, 69-81.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. *Robust estimates of location*. Princeton: Princeton University Press, 1972.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement*. Washington: American Council on Education, 1971.

Bell, A. I. *A comparison of three equating procedures on the certifying examination for primary care physician's assistants*. Paper presented at the 1979 Convention of the American Educational Research Association, San Francisco, April 1979.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley, 1968.

Bock, R. D. Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 1972, *37*, 29-51.

Bock, R. D., & Lieberman, M. C. Fitting a response model for n dichotomously scored items. *Psychometrika*, 1970, *35*, 179-197.

Fan, C. On the application of the method of absolute scaling. *Psychometrika*, 1957, *22*, 175-183.

Fleishman, A. I. A method for simulating non-normal distributions. *Psychometrika*, 1978, *43*, 521-532.

Forster, F., & Ingebo, G. *Linking groups of items*. Paper presented at the 1979 Convention of the American Educational Research Association, San Francisco, April 1979.

Fruchter, D. A., & Ree, M. J. *Development of the Armed Forces Vocational Aptitude Battery: Forms 8, 9, and 10*. AFHRL-TR-77-19, AD-A039 270. Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, March 1977.

Gugel, J. F., Schmidt, F. L., & Urry, V. W. Effectiveness of the ancillary estimation procedure. In C. L. Clark (Ed.), *Proceedings of the first conference on computerized adaptive testing* (U. S. Civil Service Commission, Personnel Research and Development Center, PS-75-6). Washington, D. C.: U. S. Government Printing Office, 1976.

Gustafsson, J. The Rasch model in vertical equating of tests: A critique of Slinde and Linn. *Journal of Educational Measurement*, 1979, *16*, 153-158.

Hambleton, R. K., & Cook, L. Latent trait models and their use in analysis of educational test data. *Journal of Educational Measurement*, 1977, *14*, 75-96.

Hughes, F. P. *The Rasch model applied to the equating of several examination forms*. Paper presented at the 1979 Convention of the American Educational Research Association, San Francisco, April 1979.

Jensema, C. J. A simple technique for estimating latent trait mental test parameters. *Educational and Psychological Measurement*, 1976, *36*, 705-715.

Kelly, P. R. *Combined common person and common item equating of medical science examinations*. Paper presented at the 1979 Convention of the American Educational Research Association, San Francisco, April 1979.

Lindgren, B. *Statistical theory*. New York: Macmillan, 1976.

Lord, F. M. A theoretical study of two-stage testing. *Psychometrika*, 1971, *36*, 227-242.

Lord, F. M. *A survey of equating methods based on item characteristic curve theory*. Research Bulletin 75-13. Princeton: Educational Testing Service, 1975.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.

Loret, P. G., Seder, A., Bianchini, J. C., & Vale, C. A. *Anchor test study final report. Project report and volumes 1 through 30*. Berkeley, CA: Educational Testing Service, 1974.

Marco, G. Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, 1977, *14*, 139-160.

Owen, R. J. A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. Journal of the American Statistical Association, 1975, 70, 351-356.

Reckase, M. D. Ability estimation and item calibration using the one and three parameter logistic models: A comparative study Research Report 77-1. Columbia: University of Missouri, Educational Psychology Department, Tailored Testing Research Laboratory, November 1977.

Reckase, M. D. Item pool construction for use with latent trait models. Paper presented at the 1979 Convention of the American Educational Research Association, San Francisco, April 1979.

Ree, M. J. Estimating item characteristic curves. AFHRL-TR-78-68, AD-A064 739. Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, November 1978.

Ree, M. J. Estimating item characteristic curves. Applied Psychological Measurement, 1979, 3, 371-385.

Ree, M. J. Personal Communication. 1980. (a)

Ree, M. J. AVRAM: Adaptive vector and response automation method. Applied Psychological Measurement, 1980, 4, 277-278. (b)

Ree, M. J., & Jensen, H. E. Item characteristic curve parameters: Effects of sample size on linear equating. AFHRL-TR-79-70, AD-A082 341. Brooks Air Force Base, TX: Personnel Research Division, Air Force Human Resources Laboratory, February 1980.

Rentz, R. R., & Bashaw, W. L. Equating reading tests with the Rasch model, Volume I Final Report. Athens, GA: University of Georgia, Educational Research Laboratory, 1975.

Rentz, R. R., & Bashaw, W. L. The National Reference Scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-179.

Samejima, F. Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 1969, No. 17.

Samejima, F. A general model for free-response data. Psychometrika Monograph Supplement, 1972, No. 18.

Samejima, F. Normal ogive model on the continuous response level in the multidimensional latent space. Psychometrika, 1974, 39, 111-121.

Schmidt, F. L., & Gugel, J. F. The Urry item parameter estimation technique: How effective? In W. A. Gorham (Chair), Computers and testing: Steps toward the inevitable conquest (PS-76-1). Washington, D. C.: U. S. Civil Service Commission, Personnel Research and Development Center, September 1976.

Slinde, J. A., & Linn, R. L. An exploration of the adequacy of the Rasch model for the problem of vertical equating. Journal of Educational Measurement, 1978, 15, 23-35.

Slinde, J. A., & Linn, R. L. A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. Journal of Educational Measurement, 1979, 16, 159-165.

Swaminathan, H., & Gifford, J. Estimation of parameters in the 3-parameter latent trait model. In D. J. Weiss (Ed.), Proceedings of the 1979 computerized adaptive testing conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1980.

Swineford, F., & Fan, C. A method of score conversion through item statistics. Psychometrika, 1957, 22, 185-188.

Sympson, J. B. The assessment of basic competencies: A new test battery. Paper presented at the 1979 Convention of the American Psychological Association, New York, August 1979.

Urry, V. W. Ancillary estimates for the item parameters of mental test models. In W. A. Gorham (Chair), Computers and testing: Steps toward the inevitable conquest (PS-76-1). Washington, D. C.: U. S. Civil Service Commission, Personnel Research and Development Center, September 1976.

Urry, V. W. ANCILLES: Item parameter estimation program with normal ogive and logistic three-parameter model options. Mimeo. Washington, D. C.: Civil Service Commission, 1978.

Vale, C. D., & Weiss D. J. A simulation study of stradaptive ability testing. Research Report 75-6. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, December 1975.

Vale, C. D., & Weiss, D. J. The stratified adaptive ability test as a tool for personnel selection and placement. TIMS Studies in the Management Sciences, 1978, 8, 135-157.

Wainer, H., & Wright, B. D. Robust estimation in the Rasch model. In D. J. Weiss (Ed.), Proceedings of the 1979 computerized adaptive testing conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1980.

Warm, T. A. A primer of item response theory. Technical Report 941078. Oklahoma City: Department of Transportation, U.S. Coast Guard Institute, October 1978.

Weiss, D. J. (Ed.) Proceedings of the 1979 computerized adaptive testing conference. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1980.

Weiss, D. J., & Betz, N. E. Ability measurement: Conventional or adaptive? Research Report 73-1. Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program, 1973.

Wood, R. L., Wingersky, M. S., & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton: Educational Testing Service, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B. D., & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

APPENDIX A--SUPPORTING TABLES

Table A-1. Characteristics of the ASVAB
General Science Subtest by AFEES

AFEES	N	Mean	SD	Skew	Kurtosis
1	500	.2759	.9975	-.2559	-.9677
3	500	-.2700	.9629	.2230	-.6493
5	500	.0424	.9233	-.2850	-.4963
6	499	.1316	1.0036	-.3345	-.5874
7	500	.1577	.9717	-.3273	-.5866
8	500	-.1189	.9899	-.0409	-.6064
9	497	-.1391	.9960	-.0434	-.7140
10	500	.0586	.9589	-.1956	-.6268
12	500	.1587	.9123	-.2064	-.7096
13	498	-.0388	.9763	-.1974	-.6223
14	498	.3436	.8849	-.4363	-.3761
15	500	-.3154	1.0679	.0466	-.7725
16	500	.0173	1.0550	-.1409	-.8760
18	498	-.3985	1.0101	.0824	-.6752
19	498	.0021	.9756	-.0912	-.8322
20	497	.4389	.8544	-.5075	-.3148
22	500	-.2880	.9980	.1660	-.7573
24	500	.1239	.9449	-.2193	-.6742
25	499	.3173	.9534	-.5289	-.4252
26	500	.2643	.9311	-.3749	-.4579
27	498	-.5292	.9194	.3814	-.4544
28	499	-.4400	.9658	.4163	-.6887
29	499	-.1850	.9564	-.0341	-.8177
30	498	-.2212	1.0073	.1015	-.7309
31	500	-.4460	.9945	.2912	-.6558
32	500	-.6476	.8614	.4003	-.1635
33	500	-.2171	1.0002	.0805	-.7691
34	499	-.0318	.9542	-.1562	-.6444
35	499	-.5602	.9253	.4211	-.3806
36	498	-.4483	.9480	.1514	-.4097
37	499	-.0875	.9508	-.1301	-.6380
38	499	-.4957	.9286	.2750	-.3721
41	500	.0943	.9005	-.1907	-.6111
42	499	.0197	.9267	-.0553	-.5823
43	499	-.1200	1.0224	-.0694	-.7847
44	499	-.0471	.8941	.0706	-.6153
45	500	-.1833	.9828	.0308	-.7571
46	500	-.2542	1.0306	.0859	-.8044
47	500	-.4734	.9692	.2842	-.4526
48	499	.0146	.9763	-.0841	-.7965

Table A-1. Characteristics of the ASVAB General Science Subtest by AFEES (Continued)

AFEES	N	Mean	SD	Skew	Kurtosis
49	498	-.1054	.8751	-.0421	-.4645
50	500	-.4349	.9739	.2451	-.8044
51	495	-.2721	.9763	.1777	-.7167
52	500	.2393	.9309	-.4302	-.3879
53	499	.3122	.9180	-.4255	-.4789
54	498	.0830	.9668	-.2421	.5689
55	500	.3658	.9486	-.5263	-.3761
56	499	.1372	.9923	-.3800	-.4164
57	500	.1026	.9327	-.0894	-.8000
58	500	.2050	1.0407	-.3811	-.6244
59	500	.3496	.9476	-.5014	-.4554
60	499	.1688	.9222	-.1803	-.7032
61	499	.3851	.9024	-.6893	.3777
62	498	-.0607	1.0162	-.1157	-.8639
63	497	.3890	.9301	-.3154	-.7974
64	500	.4154	.9066	-.4386	-.5772
65	500	.3866	.9587	-.4136	-.6086
66	500	.0442	.9446	-.0944	-.6210
67	500	-.0438	.9523	-.0587	-.6479
68	500	.1077	.9942	-.2687	-.8586
69	497	.2357	.9770	-.2619	-.8596
70	500	.4520	.8901	-.6993	.5596
71	499	.2950	.9245	-.2888	-.6018
72	500	.4413	.9064	-.6921	.2333
76	498	.2952	.9114	-.4868	-.3101

Table A-2. Items Selected for Inclusion in the Normal, Rectangular, and Peaked Anchor Tests

	True Item Parameters			Estimated Item Parameters		
Anchor Test	a	b	c	$\hat{a}$	$\hat{b}$	$\hat{c}$
Normal	2.2766	.0338	.1401	2.2717	.0078	.1059
	1.8243	-1.8344	.3763	1.4526	-2.3105	.1748
	1.7780	1.9989	.1893	3.0000	1.7863	.0955
	1.8098	.4236	.1170	2.2358	.4736	.1079
	3.8753	-.7242	.2951	3.0000	-.7405	.1901
	2.5663	-.3764	.1719	2.3082	-.4020	.0924
	1.9929	.3155	.1834	1.9821	.3446	.1689
	1.5909	1.0338	.1102	1.7310	1.1774	.1342
	2.5162	-1.1096	.1104	2.1824	-1.1509	.0059
	2.1169	-.5406	.2442	1.6920	-.6036	.1106
	2.6324	.6080	.3174	2.3324	.6768	.2907
	2.3331	.7268	.3429	1.9484	.7717	.3210
	2.1136	-1.2472	.1364	1.8686	-1.2710	.0643
	2.2304	-1.6778	.1435	2.0307	-1.5930	.1640
	2.2070	1.3933	.3067	3.0000	1.4893	.2275
	1.8899	-.0312	.1902	1.6845	-.0108	.1378
	1.8079	-.3500	.2895	1.7847	-.2940	.2531
	1.5047	-.5989	.1256	1.6149	-.4958	.1126
	1.8009	.2759	.2322	1.6597	.3591	.2240
	1.4296	.7051	.2286	1.6929	.8457	.2637
	1.7189	-.9806	.3265	1.6022	-1.0177	.2227
	1.8392	-1.5184	.1105	1.7279	-1.4533	.0377
	1.6760	1.4379	.3101	1.9381	1.5048	.3151
	1.7838	-.1039	.2143	1.4660	-.1524	.1183
	1.3747	-.4829	.1456	1.3737	-.3864	.1557
Rectangular	2.2766	.0338	.1401	2.2717	.0778	.1059
	1.8243	-1.8344	.3763	1.4526	-2.3105	.1746
	2.3086	2.1240	.1439	2.4515	2.1056	.1259
	2.0181	.9706	.1966	2.6381	.9975	.1558
	2.5162	-1.1096	.1104	2.1824	-1.1509	.0059
	3.8753	-.7242	.2951	3.0000	-.7405	.1901
	1.8098	.4236	.1170	2.2358	.4736	.1079
	2.2070	1.3933	.3067	3.0000	1.4893	.2275
	2.2304	-1.6778	.1435	2.0307	-1.5930	.1640
	2.1136	-1.2472	.1364	1.8666	-1.2710	.0643
	2.6324	.6080	.3174	2.3324	.6768	.2907
	1.6760	1.4379	.3101	1.9381	1.5048	.3151
	1.8392	-1.5184	.1105	1.7279	-1.4533	.0377

Table A-2. Items Selected for Inclusion in the Normal, Rectangular, and Peaked Anchor Tests (Continued)

Anchor Test	True Item Parameters a	True Item Parameters b	True Item Parameters c	Estimated Item Parameters â	Estimated Item Parameters b̂	Estimated Item Parameters ĉ
Rectangular (Cont.)	1.4949	-1.9274	.1493	1.2598	-2.1565	.0977
	1.3346	2.3002	.1202	2.1999	2.3542	.1633
	1.9929	.3155	.1834	1.9821	.3446	.1689
	2.5663	-.3764	.1719	2.3082	-.4020	.0924
	1.8353	-.7625	.1751	1.5589	-.8333	.0606
	2.3331	.7268	.3429	1.9484	.7717	.3210
	1.5909	1.0338	.1102	1.7310	1.1774	.1342
	1.7525	-1.8702	.2204	1.6999	-1.7462	.2693
	1.3909	-1.8081	.1144	1.3265	-1.8646	.0699
	1.3888	1.8744	.1674	1.9353	1.9720	.1973
	1.8009	.2759	.2322	1.6597	.3591	.2240
	1.5617	-.4916	.1561	1.7318	-.3962	.1286
Peaked	2.2766	.0338	.1401	2.2717	.0778	.1059
	2.5241	-.1973	.2941	2.2957	-.1850	.2327
	2.5663	-.3764	.1719	2.3082	-.4020	.0924
	2.1322	-.2409	.1218	1.8271	-.2715	.0364
	1.9838	.0308	.1765	1.8246	.0482	.1243
	2.1322	.1437	.1296	1.7626	.1053	.0788
	2.5678	-.0124	.2990	2.0325	-.0081	.2535
	1.7472	-.2444	.1108	1.7626	-.1665	.1060
	1.8899	-.0312	.1902	1.6845	-.0108	.1378
	1.8609	-.4670	.1111	1.7860	-.4194	.0573
	2.1462	-.3844	.1625	1.8270	-.4245	.0751
	2.8007	-.4404	.3155	2.4904	-.4772	.2332
	2.2596	-.0840	.2209	1.5028	-.1956	.0838
	1.5617	-.4916	.15[illegible]	1.7318	-.3962	.1286
	1.8079	-.3500	.26[illegible]	1.7847	-.2940	.2531
	2.1945	-.4158	.2529	1.7870	-.4213	.1749
	1.7838	-.1039	.2143	1.4660	-.1524	.1183
	2.1038	-.2952	.3263	1.6991	-.3497	.2141
	1.4159	-.1788	.1443	1.4204	-.1162	.1102
	1.5732	-.2968	.2128	1.5095	-.2477	.1697
	1.8253	-.1994	.2239	1.4196	-.3236	.0758
	1.9929	.3155	.1834	1.9821	.3446	.1689
	1.7777	-.3484	.2414	1.5750	-.3496	.1905
	2.2983	-.2287	.3771	1.6622	-.2980	.2831
	2.2819	-.3800	.3237	1.7573	-.4519	.2241

APPENDIX B--REVISIONS TO PROGRAM OGIVIA

The item calibration program, OGIVIA, was obtained from James McBride of the Navy Personnel Research and Development Center in San Diego. The version received was written by Jerry Edwards of the University of Washington and had been revised and updated by John F. Gugel of the U.S. Civil Service Commission. A review of the program revealed several problems. Their possible impact and the corrections made are detailed below.

A variant of the test information value was originally used for the scaling factor in the Newton-Raphson ability estimation routine. This factor was replaced with the second derivative of the log of the Bayesian posterior density function. In theory, this substitution should have made little difference in the ability and parameter estimates obtained. In fact, differences in the second and third decimal place were occasionally observed. This was assumed to be due to the fact that the criterion for termination of the iteration was a change in the absolute value of the estimate of less than 0.005 and that when the original scale factor was used, there was no assurance that the estimate was within 0.005 of the final value at this point. The differences were thus attributed to increased accuracy of estimate obtained with the modification. It was also noted that changing to the second derivative resulted in an average 20% decrease in the computer time required to estimate ability.

Another inefficiency was noted in the Newton-Raphson procedure. It appeared that this procedure, by itself, was not always successful in locating the modal Bayesian ability estimate. In some cases, the Bayesian posterior density function can be of a sufficiently irregular shape that a starting value very near the final estimate is required for convergence. The original program discarded examinees whenever the ability estimate failed to converge in 20 iterations. To preclude such examinee loss, the original algorithm was augmented by adding a bisection routine. The bisection was invoked whenever the Newton-Raphson procedure failed to converge within seven iterations. Following the bisection procedure, providing that a root existed in the interval $-8.0 \leq \hat{\theta} \leq 8.0$ (a virtual certainty), the Newton-Raphson procedure was called again to refine the estimate and was allowed to iterate up to eight times.

A final problem was encountered when OGIVIA discarded items whose parameter estimates exceeded pre-established bounds. While in practical calibration applications this may be an acceptable solution, in the present research design it presented a serious biasing effect on the comparisons of different cells in the design. To alleviate this problem, items whose parameter values would have caused them to be discarded were arbitrarily bounded as follows:

$$0.5 \le \hat{a} \le 3.0,$$

$$-3.0 \le \hat{b} \le 3.0,$$

$$0.0 \le \hat{c} \le 0.5.$$

Although somewhat arbitrary, these values appear to reflect reasonable ranges for the parameters and seemed preferable to loss of the item. These item parameters were bounded on both the first and second stages of the OGIVIA program.